R E P O R T   R E S U M E S

ED 013 523 EC 001 476

A CRITIQUE OF EXPERIMENTS ON THE ROLE OF NEUROLOGICAL ORGANIZATION IN READING PERFORMANCE.

BY- GLASS, GENE V.

PUB DATE 66

EDRS PRICE MF-$0.50 HC-$2.88 72P.

DESCRIPTORS- *READING DIFFICULTY, *NEUROLOGICAL ORGANIZATION, *RESEARCH METHODOLOGY, *READING RESEARCH, NEUROLOGICALLY HANDICAPPED, READING INSTRUCTION, METHODS RESEARCH, LITERATURE REVIEWS, CHILDREN, ADOLESCENTS, EVALUATION, C.H. DELACATO

FIFTEEN EMPIRICAL STUDIES CONCERNED WITH THE ROLE WHICH NEUROLOGICAL ORGANIZATION PLAYS IN THE TEACHING AND IMPROVEMENT OF READING ARE ANALYZED. FOLLOWING A REVIEW OF DELACATO'S THEORY OF NEUROLOGICAL ORGANIZATION, EACH OF THE STUDIES IS PRESENTED WITH ALTERNATIVE INTERPRETATIONS OF THE DATA AND WITH IMPLICATIONS NOT ACKNOWLEDGED OR CONTRARY TO THOSE DRAWN BY THE ORIGINAL AUTHORS. EACH STUDY IS ANALYZED IN DETAIL AS TO THE MANNER OF SELECTION OF SUBJECTS (THE SUBJECTS WHO PARTICIPATED IN ALMOST ALL OF THE EXPERIMENTS REPORTED IN THIS PAPER COULD NOT BE CHARACTERIZED AS SERIOUSLY NEUROLOGICALLY DISORGANIZED), THE STATISTICAL ANALYSIS OF DATA, EXPERIMENTAL TREATMENT, AND THE IMPLICATIONS DRAWN FROM THE REPORTED RESULTS. THE AUTHOR IS GENERALLY CRITICAL OF THE STUDIES FOR THEIR LACK OF ADHERENCE TO ACCEPTABLE STANDARDS FOR EMPIRICAL EXPERIMENTAL DESIGN. HIS CONCLUSION IS THAT ALL THE EMPIRICAL RESEARCH REPORTED THUS FAR HAS FAILED TO PRODUCE COGENT EVIDENCE THAT D.H. DELACATO'S THERAPY HAS AN EFFECT ON THE READING OF NORMAL SUBJECTS. IN REVIEWING STUDIES WHICH CONTAIN INFORMATION ON THE CORRELATION OF NEUROLOGICAL ORGANIZATION AND CERTAIN VARIABLES, THE AUTHOR FINDS THAT MEASURES OF NEUROLOGICAL ORGANIZATION ARE MORE HIGHLY CORRELATED WITH MEASURES OF NONVERBAL INTELLIGENCE THAN THEY ARE WITH MEASURES OF READING ACHIEVEMENT. THE FIFTEEN STUDIES ARE ALL TAKEN FROM EXPERIMENTS REPORTED IN THREE VOLUMES WRITTEN BY DECACATO AND LISTED IN THE 35-ITEM BIBLIOGRAPHY. (TM)

For ERIC Clearinghouse

A Critique of Experiments on the Role of Neurological Organization in Reading Performance

Gene V Glass *

Center for Instructional Research
and Curriculum Evaluation
University of Illin

0.0. Introduction and Outline

This is a review of fifteen empirical studies of the role which neurological organization plays in the teaching and improvement of reading. Each of the studies considered in this review comes from one of the three following sources: C. H. Delacato, The Treatment and Prevention of Reading Problems (1959); C. H. Delacato, The Diagnosis and Treatment of Speech and Reading Problems (1963); C. H. Delacato, Neurological Organization and Reading (1966). Although Delacato performed only two of these studies himself, they are all cited by him as evidence in support of his system of therapy to remediate poor readers and improve the performance of good readers through neurological training.

No new empirical evidence is presented in this paper, if "new evidence" is taken to mean "new data." However, new evidence is presented in the form of plausible alternative interpretations of the data in the fifteen published studies and new analyses which reveal implications not acknowledged or contrary to those drawn by the original authors.

The sections of this paper take the following order:

1.0 Background: Delacato's Theory of Neurological Organization and Reading

1.1 Diagnosis and Treatment of Poor Neurological Organization

2.0 General Methodological Considerations in Experiments Testing Delacato's Theory of Neurological Organization

* Now at the Laboratory of Educational Research, Uni. of Colorado.

1.0. Background: Delacato's Theory of Neurological Organization and Reading.

The central theme of Delacato's theory of neurological organization is the biology student's familiar tongue-twister "Ontogeny recapitulates phylogeny." It is Delacato's belief that the phylogenetic development of the central nervous system, which assumes its highest form in man, is reflected in the development of the nervous system of each human. It is further asserted by Delacato that if for any reason the neurological development of a child does not proceed through a certain sequence of stages, the child will exhibit difficulties in mobility and speech and in the "essence of the human nervous system, reading."

Delacato takes the position that reading difficulties stemming from poor neurological organization--the failure of a child's nervous system to develop phylogenetically--can be corrected by training the child to be neurologically well organized. This training consists of finding the stage at which impairment of the proper neurological growth took place and retraining the child from that point through the higher stages until complete neurological organization is achieved. In administering this therapy, one is supposedly "recapitulating the opportunity to develop the child's nervous system."

1.1. Diagnosis and Treatment of Poor Neurological Organization

A. Neurological organization is diagnosed at the highest, most complex level by observing whether the child has established a clear dominance of one side of the body in activities involving the feet, hands, and eyes. Mixed laterality (e.g., left-footed, right-handed, left-eyed) is evidence of poor neurological organization. High tonal ability, as indicated by an interest in music and the ability to perform musically, is also considered evidence of poor neurological organization at this level. (Many schools subscribing to Delacato's theory have eliminated music in the primary grades in the belief that it interferes with the children's attempts to learn to read.)

B. Neurological organization is evaluated at the second highest level, the cortical level, by observing whether the child walks with good balance, smoothly and rhythmically, and in a cross-pattern manner, i.e., extending right arm with left leg and *vice versa*. Smoothness of movement of the eyes during visual pursuit is taken as evidence of good neurological organization at this level.

C. At the level of the midbrain good neurological organization is indicated by smooth, rhythmical cross-pattern creeping and smooth eye movement during visual pursuit of an object held in the child's own hand.

D. At the level of pons, the lowest level evaluated, good neurological organization is indicated by a sleeping position appropriate to the child's laterality and smoothness of visual pursuit with each eye while the other eye is occluded.

The lowest level at which a child fails one or more tests is the level at which training for proper neurological organization begins. This training consists of teaching the child to perform properly those activities by which neurological organization is evaluated.

One cannot do justice to Delacato's writings on the diagnosis and treatment of neurological organization in the space available here. This section is intended only to serve as minimal information necessary to follow the discussion of the empirical studies. For a better understanding of Delacato's position, the reader must refer to Delacato (1959, 1963, 1966).

2.0. General Methodological Considerations in Experiments Testing Delacato's Theory of Neurological Organization

It is understandable that the eye of any research methodologist would be caught by a collection (Delacato, 1966) of experiments and studies which is advertised by its publisher as "the largest number of controlled scientific experiments on any single educational concept." The experiments reported in Neurological Organization and Reading are, in the opinion of their editor, conspicuous for "excellence of design and control." The book is dedicated to those "who, upon reading the original concepts of Neurological Organization, reacted as true scientists by submitting the new concept to controlled experimental scrutiny." Such a volume merits study by all educational researchers.

Unfortunately, all of the experimental studies of Delacato's theory of neurological organization and its relationship to reading are exemplary only for their faults. They were naively designed and clumsily analyzed. They suffer from a multitude of sources of invalidity. They appear to have been executed and reported in an atmosphere of relative insensitivity to basic considerations of empirical, experimental research.

A. It is amazing that research in remedial reading could be carried out and published in 1966 without the slightest appreciation of the workings of the regression effect. It is difficult to believe that research worth publishing--not to mention deserving of the judgment "a scientific appraisal of the concept of Neurological Organization"--could be executed without the least sensitivity to the inevitable fact that groups of persons chosen for the extremeness of their scores on a variable will <u>regress toward the mean</u> on subsequent observations on the same and related variables. That such a statistical-psychometric artifact would be mistaken for "gains" attributable to a therapy intended to remediate poor readers is ironic considering how prone research methodologists are to illustrate this phenomenon with examples from remedial reading research. (The following references contain discussions of the role played by regression toward the mean in many experiments: Rulon, 1941; Thorndike, 1942; Campbell and Stanley, 1963; Lord, 1963; Thorndike, 1963; Biaggio and Stanley, 1964.) Of the eleven studies in Delacato (1959, 1963, 1966) which might be regarded as experiments in which variables are manipulated, five studies are largely invalidated by the failure of the researcher to use a control group to control for the upward regression of subjects chosen because of their low scores on a pretest. In each instance, the "gain" from pretest to posttest which is attributed by the authors to the effectiveness of Delacato's theraputic techniques could have arisen easily from the natural regression toward the mean of extreme scores. This will be seen clearly when each study is discussed in detail.

B. Practically all of the studies appear to have been executed in thoroughgoing ignorance of the fundamental principles of comparative experimental design which have been known to researchers for thirty years (Fisher, 1935). Repeatedly, a "matched-groups" design is employed in instances in which the random assignment of subjects to experimental and control groups was clearly feasible. In one instance, experimental and control subjects were matched on a pretest of reading performance which was substantially different in content from the posttest which was used to assess the effects of therapy. "Matching" of experimental and control subjects on pretest measures has long been regarded by social scientists and educational research methodologists as a vestige of the unenlightened age of experimentation which preceded the contributions of Sir Ronald Fisher (Stanley and Beeman, 1958; Campbell and Stanley, 1963). Campbell and Stanley (1963, p. 185) made this point emphatically: "...while simple or stratified randomization assures unbiased assignment of experimental subjects to groups, it is less than a perfect way of assuring the initial equivalence of such groups. It is nonetheless the only way of doing so, and the essential way. This statement is made so dogmatically because of a widespread and mistaken preference in educational research over the past thirty years for equation through matching." In four of the seven comparative experiments in The Diagnosis and Treatment of Speech and Reading (1963) and Neurological Organization and Reading (1966), the matching of subjects in control and experimental groups on pretest measures was carried out in place of random assignment.

In one of the three remaining studies, it was reported that subjects enrolled "in the usual chance manner" by themselves as volunteers in the two classes--one of which was to be designated "experimental," the other "control." The experimental group met daily from 8 a.m. to 10 a.m.; the control group met from 10 a.m. to 12 noon.

C. A consistent failing of all but a single experiment in Delacato's three publications (1959, 1963, 1966) is that experimental and control pupils were treated as intact groups, not independently as separate experimental units. A valid experimental design must embody the following: (a) replications under similar conditions, i.e., more than one observation under each experimental treatment, (b) mutual independence of the replications, and (c) randomization of all uncontrolled variation in the replications (Wold, 1956). When the twenty pupils in an experimental classroom study together under condition A and the twenty pupils in the control classroom study under condition B, there exists one replication of the experiment, not twenty replications. To obtain two replications and thus the capability of performing a valid statistical analysis on the means of the classrooms, two additional classrooms--one experimental and one control--must be observed under the experimental conditions.

It was the practice in the experiments reported in Delacato (1959, 1963, 1966) to place control and experimental pupils into separate classrooms (see 3.7, 3.8, 3.11 below), often meeting at different times during the day (see 3.5 and 3.9 below), studying under different teachers (see 3.8 and 3.11 below), etc. In only one experiment (Chapter 12, Delacato, 1966) was a sufficient number of intact groups involved that a legitimate design, with random assignment of classrooms to the experimental and control conditions, could have been implemented and a legitimate analysis of the data performed. This single opportunity was wasted when the decision was made to designate all first-period classes "experimental" and all second-period classes "control."

In none of the comparative experimental studies was there evidence of awareness of the fact that when a treatment is applied to a group of subjects instead of to each subject individually and independently, an appropriate analysis of the experiment uses the means of the groups as raw data. It is not legitimate to perform the analysis on the scores of each individual in such instances. To do so is to give the impression of far greater precision in the data than actually exists. The dictum which the researcher must obey is as follows: The unit of analysis, i.e., the raw data upon which one counts up degrees of freedom, must be the same as the experimental unit, i.e., the smallest subdivision of the total group of subjects which is randomly assigned to the experimental conditions and which is treated independently of other experimental units for the duration of the experiment. It is not surprising that this dictum is consistently violated in the work which Delacato reported. An appreciation of the importance of determining the legitimate "experimental unit" and having it coincide with the "unit of statistical analysis" has not been widespread in educational research. However, the topic has received sufficient attention in recent writings on the methodology of educational research (Lindquist, 1953, pp. 192-193; Campbell and Stanley, 1963, p. 192; Lumsdaine, 1963, pp. 656-658; Page, 1965) that a total disregard of the matter can be labeled a venial sin, if not a mortal one. But if one insisted that no experimental results were worth considering unless the unit of statistical analysis and experimental unit were the same--and such insistence would not be altogether unjustified--ninety-nine percent of comparative experiments in educational research would have to be discounted. All of the comparative experiments in Delacato's three publications fall into this large class. Thus, in proceeding to discuss the comparative experiments Delacato cited as supporting his theory and therapy, one may be conceding too much in the argument over what the data actually reveal.

D. The authors of several studies could not restrain their enthusiasm for a new program designed to improve reading performance. Some of the studies which will be reviewed here are notable examples of non-objectivity in reporting findings. One cannot avoid the impression that the effectiveness of Delacato's therapy was prejudged in some instances. (Note particularly the introductory remarks to Chapter 15 in Delacato (1966) and to the study reported on pages 153-166 in Delacato (1963).)

The effects of experimenter bias are familiar to reading researchers. McDonald (1963) expressed some concern that the enthusiasm of an experimenter and the effects of a novel "experimental" atmosphere often produce a bogus (not attributable to the essential features of the new program itself) improvement (which he called the "placebo response") in reading performance:

> "Thus, placebo responses are particularly likely in reading programs where the instructors rely heavily on special instrumentation (and themselves believe in the beneficial effects of the instruments) or have found a new 'break-through method' which they believe cannot be measured by existing devices or techniques. In fact, placebo responses may account for sixty to eighty percent of all outcomes of programs which are taught by highly enthusiastic instructors who have thrown off the 'fetters' of 'old-fashioned statistical and experimental methods'.... Almost every review of research cites one or more reports of 'gains' produced by the simple device of urging the students to read faster or by the somewhat more sophisticated method of using daily rate tests without comprehension checks."

In the final study, Chapter 19, in <u>Neurological Organization and Reading</u>, Miracle acknowledged this possibility of a Hawthorne or novelty effect in his study:

"Those students who showed the greatest progress in this study (the students receiving Delacato therapy) were probably more interested in participation than were the students in Group C (one of the control groups). This interest might well account for some of their increase in reading ability. This writer feels that if such an interest did exist it was probably due to the nature of the neuro-psychological training which provides an interesting therapy for the student." (Delacato, 1966, pp. 178-179)

Miracle's small concession to those who would seek a penetrating analysis of the actual effects of Delacato's system of therapy is quite out of character with the tenor of the other research reports in the book. Generally, the possibility that the enthusiasm of the experimenter might be reflected in the results to some degree was not acknowledged.

In Chapter 15 an experimental setting which literally bristles with novelty effects is described:

"Until about Christmas we kept in touch with the control group. After the Primer Test (February 8, 1965) we went on ahead leaving the control group behind. The control group held their own all year, coming forth with a rating of 'average' in all their tests.

"After the First Reader Test (March 18, 1965), after all the basic work had been done, we [the experimental group] went on into a more individualized program. Thirty-six different titles of Dr. Suess books for children captured their attention. Also, some twenty-nine different titles of the Wonder Books Easy Reader Series and any supplementary reader borrowed from other classrooms helped us out. A Book Fair in early March was a Godsend!

"Now I had trouble keeping the children interested in their other subjects. Their style of writing became terrific." (Delacato, 1966, p. 128)

3.1. The 1959 Delacato Study

The sole empirical study, other than case studies, in The Treatment and Prevention of Reading Problems (1959, pp. 98-100) was performed on thirty pupils who showed moderate reading problems. To qualify for inclusion in this study, a pupil had to be in the lower third of his class and had to perform at least one and one-half years "below his expectancy level" on a reading achievement test. These thirty pupils were given Delacato therapy for eight weeks, after which a posttest was administered.

Delacato presented no detailed analysis of the data from this experiment. It was simply reported that the maximum, median, and minimum "gains" from a pretest to a posttest of reading performance were 2.3 years, 0.9 years, and 0.4 years in grade-placement units, respectively. The reading test used to measure reading performance was not named. Delacato presented these data as evidence for the effectiveness of his therapy. At least three sources of gain <u>other than effectiveness of the therapy</u> can be identified. First, the time elapsing from pretest to posttest was 0.2 years. An "average" group would be expected to gain 0.2 years from normal reading instruction. One might hazard the conclusion that the maximum, median, and minimum "gains" from all factors other than "normal growth" between the pretest and posttest were 2.1 years, 0.7 years, and 0.2 years, respectively. A second influence which undoubtedly produced some pretest to posttest increase--but an influence which is more difficult to evaluate without a special empirical study--is the practice effect on the posttest resulting from having taken the same test only eight weeks previously. In several studies cited by Delacato, different forms of a test or even different tests were administered as pretest and posttest. In this instance, one would like to know if Delacato administered the same form of the same test on both occasions. The third influence which can easily be mistaken for "gains" due to therapy in this crude experimental design was probably the strongest. The regression effect would be expected to produce gains from pretest to posttest which are much greater than those to be expected from both the facilitating effect of repeated testing and the effect of normal growth during the experimental period. Pupils chosen because of their extremely low scores on a pretest will "gain" on a posttest regardless of the length of time or the nature of the events which intervene between the pretest and the posttest. To fail to acknowledge and account for this phenomenon and the other influences noted above in a study in which they might well account for most of the observed "gains" renders the results of the study suspect and marks the research as naively executed.

3.2. The Piper Study

If one is to judge an author's estimate of the importance of a study by the comprehensiveness of the description he makes of it, then on pages 152 to 166 of The Diagnosis and Treatment of Speech and Reading Problems can be found an experiment which Delacato must have considered definitive.. The section beginning on page 152 is entitled "Universal Application."

The study in question was performed in 1962 by Gayle L. Piper of Mingus Union High School, Jerome, Arizona. Fourteen pupils experiencing reading difficulties were tested in February with Form 1 of the Gates Basic Reading Test. Delacato therapy was administered for six weeks at which time Form 2 of the Gates test was given. After six weeks further therapy, Form 3 of the Gates test was administered on May 1, 1962. Therapy was suspended during the summer; on September 6, 1962, Form 4 of the Gates test was given. The design of the study can be diagrammed as follows:

Test (Form 1) → Therapy → Test (Form 2) → Therapy → Test (Form 3) → Therapy → Test (Form 4)

Effectiveness of the therapy was measured by gain scores. Any difference between a test score on a child and one obtained earlier on the same child was considered evidence of gain; no attempt was made to correct this "gain" for an increase to be expected from the normal passage of time in school. Even though Delacato was careful to subtract "elapsed time" from any such "gain score" in the studies in the 1966 work, Neurological Organization and Reading, this simple precaution was not employed in Piper's study which Delacato reprinted in 1963.

The following test scores were recorded for the fourteen subjects:

Grade-placement Scores on the Gates Basic Reading Test

Pupil No.	February 1, 1962	March 15, 1962	May 1, 1962	September 6, 1962
1	3.9	4.8	4.1	4.9
2	5.8	6.9	5.2	7.0
3	3.1	4.3	4.3	5.0
4	5.7	5.7	6.2	6.9
5	6.3	6.4	6.8	8.2
6	4.3	5.2	4.5	5.0
7	5.7	5.9	5.9	7.0
8	7.0	7.8	8.3	8.7
9	4.9	5.7	5.8	6.1
10	3.4	2.9	3.1	3.6
11	3.5	3.8	3.6	Transferred
12	5.1	5.0	5.5	6.7
13	6.9	6.7	6.8	6.8
14	5.4	6.6	5.9	7.1
Mean =	5.07	5.55	5.43	6.38

At least four explanations of why posttest scores exceed pretest scores can be identified:

a. Since the subjects were chosen for therapy because of poor reading and academic performance, their scores on subsequent administrations of the reading test should increase because of the inevitable phenomenon of regression toward the mean.

b. One might expect that increases in achievement test scores would result from familiarity with the format of the test--and other factors referred to as the "practice effect of testing"--when four alternate forms of the same test are administered in a seven-months period.

c. Because different forms of the test were used on each testing occasion, any lack of total equivalence in the alternate forms of the test would be reflected in "posttest — pretest" measures of change.

d. "Posttest — Pretest" measures of gain can be expected to be positive because of the normal growth in reading performance of any pupil receiving instruction in reading.

Of these four possible influences on "Posttest — Pretest" measures of gain, (c) is least likely to have been operative in the Piper study. Influence (d) can be accounted for by subtracting "elapsed time" in grade-placement units from the gain scores. Influences (a) and (b) were most certainly present in the data.

Piper offered no statistical analyses of the data. Such analyses will be presented here. It must be kept in mind that the following tests of statistical hypotheses are merely descriptive of the variation of the data, and do not constitute tests of the scientific hypothesis that Delacato therapy is effective. The data from Piper's study are suspect at the outset in a way that no *t*-test can correct. The mean gain from pretest to posttest following six weeks of therapy is 0.48 grade-placement units; the variance of the gain scores is 0.33. This difference between pretest and posttest means is significant at the .01 level when a *t*-test was applied to the 14 gain scores. If the gain scores are corrected for elapsed time, i.e., if six weeks = 0.15 grade-placement units are subtracted from each gain score,* the mean gain is 0.33 grade-placement units. The *t*-statistic

*This technique, used repeatedly by Delacato in Neurological Organization and Reading (1966), will be viewed critically in the review of the McGrath study.

for testing the significance of this gain is $0.33/\sqrt{0.33/14}$ which is approximately 2.15; this value of t barely misses statistical significance at the .025 level with a one-tailed test. With confidence, it can be concluded that a non-random gain in scores occurred, though it remains moot whether the gain is due to regression, the practice effect of testing, or non-equivalence of the two forms of the achievement test.

The mean gain over the twelve-weeks period from the beginning to the end of therapy was 0.36 in grade-placement units; the variance of the gain scores is 0.28. The mean gain less elapsed time (twelve weeks) equals $0.36 - 0.25 = 0.11$. The t-statistic for testing the significance of the difference between .11 and zero is less than unity. One cannot conclude with any confidence that there is a significant difference between the means of the grade-placement scores in February and in May over and above the expected gain due to elapsed time.

The mean gain from the pretest in February to the final posttest in September was 1.19 grade-placement units. This gain is statistically significant even when elapsed time is subtracted from the gain scores. (It must be pointed out, however, that the elapsed-time score does not include two months of the summer recess. Thus, if some or all of the subjects received instruction in reading during the summer, the elapsed-time score used to correct the gain scores in the analysis would have been too small. Piper does not report whether the subjects were instructed in reading during the summer recess.) The significance test, it must be remembered, reflects only on the reliability of the gain; it does not reflect on the cause of the gain. The gain from February to May is smaller than the gain from February to September even when corrected for elapsed time. This is to be expected from data whose movement is governed by the regression effect. Since Form 1 in February correlates higher with Form 3 in May than it does with Form 4 in September, the regression toward the mean will be greater when measured from February to September than from February to May.

Piper's study does not constitute a controlled evaluation of the Delacato therapy. No cognizance was taken of the effects of regression toward the mean and practice on the achievement tests. In addition to these most important features, the "gain" from beginning to end of therapy was far from statistical significance (t was less than 1.0) when "gains" were corrected for elapsed time. The Piper study is simply not admissible evidence on the question of the effectiveness of Delacato's therapy.

3.3. The 1963 Delacato Study

On pages 170-173 of The Diagnosis and Treatment of Speech and Reading Problems (1963), Delacato reported on an experiment of his own design and execution which is purported to show that therapy to enhance neurological organization can improve performance on tests of verbal aptitude. The 25 members of the junior class of a private school for boys were the subjects in this experiment. As part of the College Entrance Examination Board's testing program, all of the 25 subjects took the Scholastic Aptitude Test. Their scores on the Verbal subtest of the SAT were recorded. At this point, Delacato formed an experimental and a control group. There is only about one legitimate way in which to do this: randomly assign some number of the subjects to the group which will receive neurological training and place the remaining subjects in the control group. There are numerous ways to form the two groups so that unknown amounts of bias result: (a) try to "match" subjects in predesignated control and experimental groups, (b) ask for volunteers for the experimental group, (c) let the boys who play football be the control group, etc. Of all the incorrect ways of forming the two groups, Delacato chose, perhaps unwittingly, the poorest and the one which was most biased in his favor. The nine lowest scoring subjects on the pretest with the SAT-Verbal were designated the "experimental group"; the 16 highest scoring subjects, the "control group." As can be seen by anyone who understands the rudiments of statistical regression, this "design" is certain to show "gains" for the experimental group and much smaller "gains" or even losses for the control group irrespective of the effectiveness of the treatment given to experimental subjects.

Neurological training was administered to the nine experimental subjects for one-half hour each day for six weeks. At some unspecified time following the end of the six-weeks experimental period, the Scholastic Aptitude Test was administered a second time. The following results were obtained:

	Pretest Mean	Posttest Mean	Mean "Gain"
Experimental Group (n=9)	547.4	554.2	6.8
Control Group (n=16)	397.7	463.5	65.8

It should come as a surprise to no one that the mean "gain" for the experimental group was considerably larger than the mean "gain" for the control group (65.8 points versus 6.8 points). How might one account for the large "gain" of the experimental group? How can one explain the fact that the control group actually gained from pretest to posttest instead of regressing downward?

The answer to the second question is simple. The regression effect does not imply that a group chosen because of their high scores at Time 1 will have a smaller mean at Time 2. This will only be necessarily true if the mean and variance of the total group from which the "high group" was selected do not change from Time 1 to Time 2. If both sets of scores (Time 1 and Time 2) are standardized to the same mean and variance, the "high group" at Time 1 will yield a lower mean at Time 2. In Delacato's study the mean of the total group of 25 scores increased 28 points from the pretest to the posttest, probably as a result of maturation of the subjects and the practice effect of having taken the SAT-Verbal test once before. Both maturation and the practice effect of pretesting operated to increase the scores of both the control and experimental groups. The regression effect joined these two influences to produce a large "gain" for the experimental group; in the control group, it militated against these influences which were strong enough, nonetheless, to produce a pretest to posttest gain where it might not have been expected.

Delacato made an attempt to provide somewhat better control for this experiment than his first control group of 16 students by going into the records of the Junior class of the previous year to measure the gain made by the nine lowest scorers on the SAT-Verbal test from a first to a second testing. With this "control group" the interval between the pretest and posttest was seven months. Since Delacato did not report the time interval between the pretest and the posttest for his experimental group, the appropriateness of the improvised control group is questioned. This consideration aside, the results Delacato obtained on these nine students are nothing short of amazing. Not only did they not regress upwards--they were the nine lowest scorers on the SAT-Verbal for the previous Junior class--not only did they fail to show a normal gain in verbal skills due to growth during the year, not only did they show no gain due to the practice effect of taking the test, but they actually showed an average <u>loss</u> of 19 points from the pretest to the posttest. This result is known to be so atypical that this improvised "control group" is without question inappropriate. Far better control would have been obtained--and quite likely a sizable gain reflecting regression, maturation and the practice of testing would have been found--if Delacato had reported the SAT-Mathematical Scores of his original group of nine experimental subjects.

3.4. The Sister M. Edwin Study

The first experimental study in Neurological Organization and Reading (1966) is reported on pages 50-53 of Chapter 8. This study was conducted by the Archdiocesan Reading Service of Chicago by Sister M. Edwin, S.C.C. A total of 108 kindergarten children began the study; 84 children "were able to participate in the total program to the end of the study." Thus, the "mortality rate" was 22 percent. The experimental period was six weeks. Subjects were pretested on June 24 with the Harrison-Stroud Reading Readiness Test. Of the 84 children who persisted through the study, 43 were in the experimental group and 41 were in the control group.

The experimental group was placed on a daily 80-minute regimen of neurological training. The control group participated in coloring, games, lunch, and a rest period, but was given no neurological training. Oddly enough, the experimental group listened to stories, folk songs, and nursery rhymes for 25 minutes each day. The only possibility for the control group to receive a comparable activity was provided by the teacher asking the mothers or some older members of the families of the control children to read or tell a story to the child for at least 10 minutes each day. Even if the families of the control subjects followed this suggestion faithfully--which is difficult to imagine--each experimental child would still have been exposed to approximately five or six more hours of such activities than each control child. It is difficult to conceive why the researchers allowed this factor of exposure to the reading and reciting of material to differ so greatly from experimental to control group.

After six weeks of neurological training, posttests were administered and the control and experimental groups were compared. The 43 experimental pupils and the 41 control pupils who persisted throughout the experiment were said to be matched on age, sex, and "knowledge of the ABC's." This assertion by the author of the research report is difficult tc confirm. No data are given to support it; nor is it clear how an exact matching of the 43 experimental and 41 control subjects was possible when experimental and control groups had to be designated at the outset of the experiment and some 24 pupils dropped out of the program. If the original 108 pupils were split into matched experimental and control groups, then the final experimental and control groups would be "matched" only if the same types of pupils dropped out of both groups. But this equivalence of the "drop-outs" would be unusual. One might expect a greater number of "drop-outs" from the experimental group or at least a different type of "drop-out" because of the rigorous experimental regimen. Parents who were generally unsympathetic to the procedures of the school might not have discouraged their child from quitting the program or might even have encouraged it when they learned of the strange approach being taken to ready their children for reading instruction. Similar unsympathetic parents of children in the control group would find nothing objectionable in the rather prosaic climate of the control kindergarten, and consequently their children would be more likely to complete the six-weeks period. In short, when 22 percent of the subjects who began the experiment did not persist, a detailed explanation of the frequency and characteristics of "drop-outs" from both the experimental and control groups is in order.

Granting that the experimental and control pupils were matched on age sex, and knowledge of ABC's--which is difficult to believe considering the above observations--were the two groups comparable for the purpose of evaluating the effect of Delacato therapy on reading readiness? It is impossible to know whether they were comparable. Surely matching on age, sex, and knowledge of ABC's does not ensure that the control and experimental groups were comparable in reading readiness to begin with. Although the pupils were pretested on reading readiness, no data of any sort are given to indicate comparability of the two groups on this variable.

The only workable and valid solution to the problem of determining comparable experimental and control groups at the beginning of the study would have been to assign pupils to either group strictly at random. Appropriate statistical tests applied to the posttest data would have answered the question whether final differences between the control and experimental groups could be attributed to their initial chance non-comparability on any and all variables with respect to which the pupils could be measured. Random assignment of subjects to the experimental and control groups was not used in this study. Indeed, no clear indication of how subjects were designated "experimental" and "control" is given. Thus, one can question the initial comparability of the two groups with respect to reading readiness and the potential to acquire reading readiness by means other than neurological training during the six weeks experimental period.

Little data and almost no statistical analysis of the objective results of this study were given. It was reported that in the experimental group the "percent of increase in score at the time of the second test averaged 82.4 percent." The comparable percent--and it is unclear what this percent means or how it was calculated--was 37.2 for the control group. Does this statement mean that 82.4 percent of the experimental subjects made a higher score on the posttest than on the pretest. Or would the author of the report have considered a gain from a score of 100 to one of 182.4 a "gain of 82.4 percent"? The author's intended meaning was not clear.

The following data have been summarized from the research report and statistical significance tests have been performed:

Variable and its Measurement	Experimental Group		Control Group		z-test of Differences Between Independent Percents
	Frequency	Percent	Frequency	Percent	
Gain in Controlled Attention Span	34	79.1	20	48.7	z = 2.7
Gain in Uncontrolled Attention Span	30	69.7	24	58.5	z is less than 1.0.
Gain from Reading Readiness Category Type 5 to Type 4	7	16.2	3	7.3	z = 1.2

The above table shows that there was a significantly larger proportion of subjects making gains on a test of controlled attention span in the experimental group than in the control group. The differences between the two groups on the other two variables were not statistically significant.

The statement in the research report that the experimental group had better than a 200 percent advantage over the control group in moving from category Type 5--the level at which pupils are considered not ready for first-grade training--to Category Type 4 is simply fatuous. Such a statement is not altogether unlike claiming that coin A is more than twice as apt to turn up "heads" than coin B, because in these 10 tosses of both coins, A yielded 7 heads and B yielded only 3.

Thus we see that the comparability on relevant variables of the experimental and control groups at the outset of the experiment is questionable, the two groups were not treated equivalently in all respects (the experimental group listened to reading and recitation for 25 minutes each day), and two of the four differences reported between the two groups did not approach statistical significance (insufficient data were reported to allow a test of one of the four differences).

3.5. The Masterman Study

In Chapter 12 of Delacato (1966) an experiment performed by Masterman is reported. A group of 422 children at two separate summer reading centers was involved in a six-weeks experiment comparing pupils receiving neurological training with a control group. The pupils were between 7 and 13 years of age; all pupils had normal intelligence but evidenced some reading problems. The 422 subjects were assigned to experimental and control groups in some unspecified manner. Nineteen teachers participated in the experiment; each teacher taught two classes. <u>Arbitrarily, each teacher's first-period class was designated "experimental" and the second-period class "control</u>." We see, then, that if there is any advantage to studying in the first-period as opposed to the second-period class of a teacher, this advantage would favor the experimental group in Masterman's study. This confounding of the effect of "time-of-day" with the experimental treatment was a major oversight. It could not be controlled by matching the experimental and control groups on pretests of reading achievement, as was eventually done, because pretesting came before the summer session. This flaw in the design could have been easily remedied by letting the experimental group be the first-period class for about half of the teachers, chosen at random, and letting it be the second-period class for the remaining half.

Subjects in the 19 experimental classes were required to report to class 15 minutes early and stay 15 minutes after class for neurological training. Of the 422 subjects who started the study, only 282 persevered or were retained through the six-weeks course, the posttest and the final matching of control and experimental subjects. Thus, the mortality rate was 33.2 percent. From the 282 subjects who completed the study, 141 pairs containing one experimental and one control subject each were formed so that pair-mates matched on sex, age and grade placement.

The data (Delacato, 1966, p. 113) show the experimental and control groups perfectly matched on pretest scores on the _Gray Oral Reading Paragraphs Test_, each with a mean of 3.80 yrs. On the posttest with the Gray Test, the 141 experimental subjects scored 4.36 yrs., and the 141 control subjects scored 4.22 yrs. (The variance of the "gain" scores in each group was 0.36.) A statistical hypothesis test (correlated _t_-test) of the difference between the control and experimental group means gave a _t_-value of 2.46 with 141 degrees of freedom. Hence, the difference was significant. Oddly enough, Delacato took these same data and performed an inappropriate independent groups critical ratio test on them in a footnote on page 114. This was clearly inappropriate and unnecessary. What is even more puzzling is that no statistical significance test was offered by either Masterman or Delacato of the _Stanford Reading Achievement Test_ data which were also gathered. It is simply reported that the mean "gain" of the experimental group from pretest (Form L) to posttest (Form M) was 0.35 yrs. for the experimental group and 0.12 yrs. for the control group. No variances were given, so it is impossible to test the reported data for statistical significance.

Actually, since classes were treated during the experiment as intact groups, the experimental unit was the "classroom" and not each individual pupil.

Hence, the analysis should have been performed on the 38 classroom means instead of the 282 individual scores. Assuming that there were about eight students per class on whom complete data were available, one can reconstruct from the data reported an <u>uncorrelated</u> *t*-test using the classroom means as the unit of analysis. This analysis shows statistical significance favoring the experimental group at the .05 level ($t_{36} = 2.0$). The most legitimate analysis would have to take into consideration the dependencies in the classroom means induced by the fact that each teacher taught one experimental and one control class. This legitimate analysis which cannot be reconstructed from the reported data undoubtedly would have shown statistical significance also.

Of the 422 subjects who began the experiment, 282 persevered to the end. It is not clear from Masterman's report whether the matching of subjects into matched pairs took place before or after the experiment. We shall assume that the determination of pairs matched with regard to sex, age, and pretest score on the Gray Test took place before the experiment. Might the fact that 33 percent of the matched pairs dropped out of the study indicate a possible bias in the comparison of the control and experimental groups? This seems quite probable. As Masterman reported, the means on the pretest for the experimental and control groups for the 141 matched pairs staying in the study were identical (3.80 yrs.). It is obvious, however, that both members of each pair were not matched on all relevant variables. One might ask: How did the 141 experimental and 141 control subjects compare on their desire to learn, general mental ability, perseverance, or interest in school work? The answers to these questions are particularly important because of the high mortality rate of subjects. If the therapy given to the experimental group was taxing or quite demanding (recall that experimental subjects were required to report to class 15 minutes early and stay 15 minutes late), then one might reasonably expect that poorly motivated, disinterested subjects would tend to drop out of the program or be absent from school at a higher rate than motivated and interested subjects. The fact that the experimental group classrooms met earlier in the day might also be related to the subjects reasons for leaving the program. Masterman was careful to eliminate a pupil's matched pair from the study whenever one pupil "dropped-out." However, since the experimental and control subjects were not matched with respect to motivation, the surviving experimental subjects would be expected to be more highly motivated on the average than their matched control subjects. In a real sense, the bias thus introduced by differential reasons for leaving the experiment for the experimental and control groups is another manifestation of the phenomenon of regression toward the mean. We can consider each matched pair as a unit. The X variable observed on this unit is the experimental subject's motivation. A demanding experimental group treatment (as in the Masterman study) would "select" and eliminate the lowest scorers on the motivation variable among the experimental subjects. Since the correlation of X and Y over the matched pairs is far from perfect, the control subjects who are paired with the low scoring experimental subjects will tend to have scores on Y nearer the mean of all Y scores. Consequently, low scoring experimental subjects who "drop-out" take with them matched control subjects who score systematically higher on motivational variables.

Evidence pertaining to possible biases of this nature in the Masterman study could have been obtained by noting which member of each matched pair--the control or the experimental subject--initiated the "dropping out." No data of this sort were presented in the research report. Fortunately, the report of an experiment by Glaeser in Chapter 16 of Delacato (1966) was sufficiently detailed that the speculations in the above paragraph concerning explanations of "mortality" in the experiment can be given some empirical support. Glaeser's experiment was quite similar to Masterman's in that the experimental group met from 8 AM to 10 AM each morning and the control group met from 10 AM until 12 noon. Of the three subjects who chose to quit coming to the experimental group, two of them gave as their reason that the class was "too early" and that it was "too far to walk to school." None of the five subjects who dropped out of the control group gave a reason which could be interpreted as low motivation to remain in class. These facts reinforce our suspicion that the time of day at which the experimental classes met and the greater demands placed on the pupils in these classes may have resulted in differential mortality between the experimental and control groups in Masterman's study. At best, the results of Masterman's experiment are inconclusive and must be regarded cautiously until additional data are published.

3.6. The McGrath Study

In Chapter 13 (Delacato, 1966) Father Francis McGrath reported on a study performed in a summer remedial reading program. Ninety-two pupils, ranging in age from approximately 12 to 16 years, were tested at the beginning of the summer on Form Am of the Metropolitan Reading Test. All 92 students were reading below their grade level. Having been chosen for their poor performance on the reading test at the outset of the experiment, one would expect them to regress upwards toward the means of the groups from which they were selected on subsequent testings. For six weeks, neurological training was administered in the form of cross-pattern creeping and walking, homolateral patterning, and attempts to establish hemispheric dominance by blocking vision with the subdominant eye. At the end of the six-weeks experimental period, an alternate form, Form Bm of the Metropolitan Reading Test, was administered.

Father McGrath presented no statistical analysis of the data. Analysis was provided in a footnote by Delacato on page 117. The average gain from pretest to posttest for the 92 subjects was 0.63 grade-placement units. Since the elapsed time (six weeks) equaled 0.14 grade-placement units, Delacato ran a correlated *t*-test on the pretest to posttest gains and obtained a "critical ratio"--actually, a *t*-statistic with 91 degrees of freedom--for testing the hypothesis that the difference between the pretest and posttest population means was 0.14. (Delacato's calculations are slightly in error. His estimate of the variance error of the difference scores was biased because he divided by n instead of $n - 1$; however, this is an inconsequential error.) The value of the *t*-statistic was 5.10.

In actuality, Delacato tested an irrelevant hypothesis. One can identify at least two influences other than the simple passage of instructional time (six weeks or 0.14 yrs.) which would cause a pretest to posttest "gain." As mentioned earlier, the regression effect was undoubtedly operative in this experiment. Indeed it may account for the major portion of the observed "gain" in scores over the six-weeks period. McGrath made no attempt to control this influence either by forming a control group or by estimating the expected increase form pretest to posttest due to regression. The second influence which was probably operative to a lesser extent than the regression effect was the practice effect on the posttest of having taken an alternate form of the test only six weeks previously. That any sizable portion of the pretest to posttest gain could be attributed to non-equivalence of Forms Am and Bm of the *Metropolitan Reading Test* is considered only a remote possibility and not a major criticism. Thus, given the experimental design and knowledge of how the experiment was carried out, one would expect a far greater increase in scores from pretest to posttest than that attributable to the passage of time alone, viz., 0.14 yrs.

Even the hypothesis, implicit in Delacato's statistical test, that this group should have shown an increase of 0.14 grade-placement units from pretest to posttest on the Metropolitan in the absence of effective therapy is highly questionable. On the one hand, since these pupils were below average for their grade and in need of remediation, they might not be expected to show a normal growth in six-weeks time which is equivalent to six-weeks growth in a normative population. On the other hand, the six-weeks of instruction in reading during the summer might be more effective than six-weeks instruction during the school year for the normative population, because of the momentum generated by nine-months of instruction during the school year, because no other courses competed for the pupils' attention, and because two periods instead of one were spent in reading instruction each day. How much progress would a comparable group of readers, which did not receive neurological training, make under these circumstances? This question could only have been answered in the McGrath study by the inclusion of a control group.

The necessity of a control group in this situation is emphasized by consideration of a study reported in Chapter 18 in Delacato (1966). A group of 40 control pupils, ranging in age from 6 to 14 years and with reading performance below grade level, showed a gain on the Stanford Reading Achievement Test of 0.40 years in a six-weeks summer remedial reading program. These pupils did not receive neurological training. In Chapter 12 of Delacato (1966), a group of 141 control subjects was given the Gray Oral Reading Paragraphs Test immediately before and after a six-weeks summer session. The average gain for the group in grade-placement units was 0.42 years. In Chapter 14 of Delacato (1966), Kabot's control group of 96 subjects showed a mean gain of 0.60 yrs. in an eight-weeks study. While these studies do not indicate that a "natural" gain of about 0.40

years should have been expected for the six-weeks period in McGrath's study, they do highlight the dangers of assuming without evidence (as Delacato did) that a non-existent control group should show an increase in reading achievement of 0.14 years in the six-weeks summer session.

Consideration of the three influences present in the data of the McGrath study renders the results and conclusion of the study suspect. The study should probably not be regarded as supporting Delacato's claims.

3.7. The Noonan Study

In Chapter 17 (Delacato, 1966) a study by Noonan is reported which is practically identical in design, results, and analysis to that performed by McGrath. Nine retarded readers in the sixth and seventh grades were placed in an experimental group to test the effects of Delacato therapy. Each subject was tested in September on Form Am of the *Iowa Silent Reading Test*. Two subjects entered the experimental group in the second semester. Form Bm of the *Iowa Silent Reading Test* was administered in June. During the entire year, 45 minutes of neurological training was given each day.

Noonan reported only the mean differences in grade-placement scores from September to June for the eleven subjects: (1) Reading rate: gain of 3.3 yrs.; (2) Comprehension: gain of 3.1 yrs.; (3) Directed Reading: gain of 3.2 yrs.; (4) Word Meaning: gain of 1.4 yrs.; (5) Paragraph Comprehension: gain of 3.7 yrs.; (6) Sentence Meaning: gain of 1.2 yrs.

The same influences which invalidated the McGrath study are present in this experiment: scores would increase from September to June because of the phenomenon of regression toward the mean; scores might increase somewhat because of a practice effect on the test; the same form of the test was not used in both September and June.

The last influence can be discounted wholly in Noonan's study. Forms Am and Bm of the Iowa Silent Reading Tests: New Edition appear to be equivalent in all important respects. However as a general research strategy, different forms of a test should not be given before and after an experiment. It is a simple matter to counter-balance the tests by giving Form A to a randomly chosen half of the subjects at Time 1 and Form B to the same subjects at Time 2; the other random half of the group receives Form B at Time 1 and Form A at Time 2.

The practice effect of having taken an alternate form of the test nine months previously is probably negligible. Hence, the "gains" which Noonan shows cannot be attributed to any practice effect of testing. However, the regression effect is a different matter. As in the McGrath study, the experimental subjects were chosen because of their poor performance on a pre-experimental measure of performance. Naturally they would not be expected to perform as poorly on a subsequent test of reading performance because of regression toward the mean.

Noonan offered no statistical analysis of his data. Analyses were provided by Delacato in a footnote on pages 147-149. A correlated *t*-test was used to assess the significance of the difference from zero of the mean "gain" score. A "gain" score equaled (Posttest grade placement) — (Pretest grade placement) — (Elapsed time). The "Elapsed time" was taken to be 0.88 yrs., i.e., 1.0 years for nine subjects, 0.5 years for the tenth subject, and 0.2 years for the eleventh subject. (Although subjects #10 and #11 participated in the therapy only 6 months and 2 months, respectively, there is some indication in Noonan's report that they were pretested in September which would imply that the "elapsed time" should have been 1.0 years for all eleven subjects.) The assumption that the experimental subjects would be expected to show a normative true growth (apart from the regression effect) of 1.0 years during the school year was questioned in connection with the McGrath study in Section 3.6. As Noonan pointed out parenthetically, seven of the eleven experimental subjects participated in a remedial reading class the previous year and made gains of from 1.0 to 1.5 years in reading achievement. This fact casts doubt on the assumption that the experimental group of eleven subjects should show a normal rate of growth of 1.0 years during the experimental period and as a consequence makes the "elapsed time" correction of 0.88 yrs. somewhat dubious.

Setting aside for the moment the question of the inappropriateness of the design and, hence the data, for evaluating the effectiveness of the therapy, let us look at the results of Delacato's statistical analysis. The values of the *t*-statistics reported by Delacato for the mean Pretest–Posttest–Elapsed Time scores for the six subtests were as follows:

(1) Reading Rate: $t = 3.68$; (2) Comprehension: $t = 3.10$;
(3) Directed Reading: $t = 6.29$; (4) Word Meaning: $t = 1.14$;
(5) Paragraph Comprehension: $t = 3.21$; (6) Sentence Meaning: $t = 0.75$.

A large computational error in the calculation of the mean Posttest–Pretest–Elapsed Time score for Paragraph Comprehension was made. Delacato reported a value of 1.7 years (p. 149); the correct value is 2.8 years. One hopes that the other calculations reported in the editor's footnotes throughout the book were made more carefully.

The *t*-tests were presented here for descriptive purposes only. They are tests of an irrelevant statistical hypothesis. Without a randomly comparable control group there is no way of assessing the proper constant by which the Posttest — Pretest gain scores should be corrected so that they reflect only improved reading performance due to neurological training. Simple statistical manipulations will not overcome faults in the design out of which the data were gathered.

3.8. The Kabot Study

In Chapter 14 of Delacato (1966) an experiment performed by Ruth Rader Kabot is reported. A control versus experimental group design was employed with both pretesting and posttesting. Kabot reported only that the Stanford Reading Achievement Test was used as a pretest and the California Reading Test as a posttest. Presumably these tests are the Stanford Achievement Test: Elementary Reading Test and the Reading section of the California Achievement Tests-Primary Level.

Ninety-six experimental and ninety-six control subjects were matched with respect to reading achievement on the Stanford Reading Test. Matching was also performed on Kuhlmann-Anderson IQ, "reading retardation, and laterality." The ninety-six experimental subjects received remedial reading instruction and exercises "advocated by Dr. Delacato for building body balance and laterality." The control group received only remedial reading instruction. The duration of the experimental period was eight weeks. Kabot gave no indication whether the control and experimental groups had the same or different teachers, met at the same or different times during the day, etc. Posttest observations were made and gains were calculated using scores on the Stanford Reading Test as the pretest and the California Reading Test as the posttest. A logical question to ask is "Were the groups initially matched with respect to the California Reading Test?" By no means can it be confidently answered "yes." If one inspects the two tests in question, one finds that they are substantially different in content. The reading section of the Stanford Test comprises 50 items in which sentences in context must be completed by choosing the appropriate word from among four options. Only 20 of the 85 items of the California Reading Test are of this sort. The remaining 65 items of the California Test involve recognizing synonyms and antonyms (40 items), identifying spoken words (15 items) and following directions (10 items). No evidence was given in the research report that the two groups were comparable on the California Reading Test at the beginning of the experiment.

The following data were reported by Kabot:

	Mean IQ	Stanford Test	California Test	Mean Improvement
Experimental Group	96	2.3	3.1	.8
Control Group	96	2.3	2.9	.6

Although Kabot reported no statistical analyses of the data, Delacato analyzed them in a footnote on pp. 120-121. A correlated t-test, employing the differences between matched-pairs gain scores, was run to test the hypothesis that the population mean of such scores was zero. There were 13 matched pairs of subjects at the beginning of the study; two of these pairs were dropped from the study when one member of the pair transferred out of school during the eight weeks of therapy. The t-statistic for the correlated t-test run on the 11 pairs taking the California Reading Test immediately after therapy equaled 1.54, which does not exceed the 95th percentile in the t-distribution with 10 degrees of freedom.

Delacato reported in the same footnote that retests (of an unspecified type) were given one year after therapy. He reported that a difference in gain of 0.54 yrs. favoring the experimental group over the control group was obtained.

Without explanation, however, the sample size shrunk from 11 to 7 matched pairs. The $\underline{t}$-statistic for these seven matched pairs and an average difference in gain scores of 0.54 yrs. is 2.84, which is significant. However, this result stands in need of elaboration: why were four matched pairs dropped and what test was given one year after treatment? Again the high mortality of subjects raises the concern for comparability of the control and experimental groups on variables related to any possible differential mortality factors which have not been "matched out" but which might be acting on the two groups.

3.9. The Sister M. Vivian Study

Chapter 15 in Delacato (1966) is the report of a comparative experiment carried out by Sister M. Vivian. In September 1964 a group of 90 first-grade pupils was divided in some unspecified manner into a control and experimental group of 45 subjects each. The median Kuhlmann-Anderson IQ scores for the experimental and control groups were 114 and 115, respectively. The experimental group earned an "average" score on a reading readiness test given on October 20; the control group scored "high average." It was reported that at the outset of the experimental period the two groups were "about as equal as they could be." The experimental group performed Delacato exercises for thirty minutes each day.

A test given to the control and experimental groups on December 9, 1964 showed that "we [the experimental group] had surged ahead [of the control group]." This is a remarkable finding--if the "surging ahead" can be attributed to the Delacato exercises--since at this date the experimental subjects had received only about 17 hours of therapy.

On the posttest of reading performance (Bond-Clymer-Hoyt Developmental Test) given in early May, 1965, the experimental group had a mean of 3.45 yrs. and a standard deviation of 0.59 yrs. The mean and standard deviation of the control group on the same posttest were 2.92 yrs. and 0.67 yrs., respectively. The difference between means was statistically significant. Forty-two of the forty-five original control subjects completed the experiment; only one of the original forty-five experimental subjects was lost. No explanation of the loss of data for three control subjects was offered.

This experiment was cited twice in Section 2.0 of this paper as an example of possible bias due to novelty, interest, and motivational effects generated by the obvious enthusiasm of the experimenter. Portions of the research report which are pertinent in this regard can be found in Section 2.0 and on p.127 of Delacato (1966).

3.10. The Glaeser Study

Chapter 16 of Delacato (1966) is a report of a comparative experiment performed by George Glaeser with the assistance of Sandia DeWaide and Rosalie Levi of Mt. Miguel High School, Grossmont, California. In the summer of 1964, sixty-six students volunteered for two reading clinic classes. Each class met for two hours daily for seven weeks.

The class of 30 students which met from 8 a.m. to 10 a.m. each morning was arbitrarily designated the experimental group. The second-period class of 36 students which met daily from 10 a.m. until 12 noon was designated the control group. Obviously, as in several other of the experiments reviewed, the effect of "time of day" is confounded with the effect of the experimental treatment. Nor can one have much faith that the control and experimental groups were reasonably comparable at the outset of the experiment. In discussing how the assignment of a student to either the control or experimental group took place, the authors claimed only that the students enrolled in either the first or second class of the day "in the usual chance manner." Undoubtedly, there is very little that is random about a student's decision to sign up for the 8 a.m. class instead of the 10 a.m. class. Although pretest data in the form of subtest scores on the Stanford Achievement Test were available, no statistics comparing the experimental and control groups at the start of the experiment were published in the report.

Both the experimental and control groups received reading instruction for seven weeks. For the experimental group, one hour of each two-hour class period was spent in a wide variety of exercises designed to improve neurological organization. At the end of the seven-weeks experimental period, the experimental and control groups were compared in terms of pretest to posttest "gains" on seven subtests of the Stanford Achievement Test (Form L). The analyses of the data in Tables I and II on page 140 of Delacato (1966) showed significantly greater gains for 15 experimental subjects than for 24 control subjects on two of the seven subtests: Paragraph Meaning and Word Meaning. No significant differences in average gain for the experimental and control groups were found on the Spelling, Language, Arithmetic Reasoning, Arithmetic Computation, and Study Skills subtests.

It will be argued that the execution of the experiment and the analysis and reporting of the data were so dubious that Glaeser's results must be discounted either until the study is replicated or more data are published. It is impossible to account for the dropping of several subjects from the experimental group prior to analysis of the data, to justify the switching of subjects between the experimental and control, or to measure the influence on the results of the selective mortality of subjects during the experimental period.

At the outset of the experiment, the control group contained 36 subjects and the experimental group contained 30 subjects. Posttest data from which "gains" were calculated and on which the two groups were compared was reported for only about 14 experimental subjects and 24 control subjects. As nearly as can be determined from the research report, the following accounts can be made of only 8 of 16 subjects lost from the experimental group and 8 of the 12 subjects lost from the control group:

Experimental Group

1. One pupil (male) transferred to a speed reading class. This represents the loss of a capable student from the experimental group and thus might have biased the study against showing greater improvement in reading performance for the experimental group.
2. Two boys transferred to a review class in mathematics. This represents a possible culling of two less capable subjects and thus may have biased the experiment in favor of the experimental group.
3. Two boys were transferred to the <u>control group</u> because of a "program change in math." If the "program change" was to provide remedial math instruction, these transfers might have amounted to taking two less capable pupils <u>out</u> of the experimental group and placing them <u>into</u> the control group.
4. One boy dropped out because of a severe sunburn.
5. Two boys dropped out because the experimental class met "too early" and it was "too far to walk to school." This represents a loss from the experimental group of two poorly motivated pupils. Since the control group met from 10 a.m. to 12 noon, poorly motivated pupils would not be expected to drop out of the experiment because school began "too early."

The above exhaust the reported accounts of losses of subjects from the experimental group.

Control Group

1. Two girls were transferred to a speed reading class. This represents the loss of two capable readers from the control group. One capable pupil (male) was lost from the experimental group for the same reason. A slight bias in favor of the experimental group might have resulted.
2. One boy was transferred to an art class. Does the substitution of an art class for a remedial reading class indicate that the pupil's reading performance was improved? If so, this transfer would have lowered the net intellectual assets of the control group.
3. Two girls dropped due to parent's illness.
4. Two girls dropped due to a change of vacation plans.
5. One girl dropped for an unknown reason.

The loss of seven of the original nine girls in the control group left two girls in the control group compared with four in the experimental group. Moreover, a greater proportion of girls existed in the experimental group than in the control group. It is a thoroughly documented fact that girls' achievement and progress in reading is superior to that of boys.

To summarize, the above reported losses appear to have resulted in biases in favor of the experimental group in terms of motivation, number of girls, and net intellectual assets.

Accounts were given of the loss of only eight pupils from both the experimental group and the control group. Given the initial figures of 36 in the control group and 30 in the experimental group, one would expect that final analyses of the data from the experiment would be based on 28 control subjects and 22 experimental subjects. However, unaccountably, only about 14 of the experimental subjects and 24 of the control subjects were represented in the data analysis. No account of why the data for approximately 8 experimental subjects and 4 control subjects do not appear was offered.

Any comparison of "gains" in reading achievement during the seven-weeks experimental period is uninterpretable in light of the fact that data for half of the experimental subjects and a third of the control subjects were either lost for differential reasons or unaccountably missing in the final analysis. The two groups of subjects were not equivalent (randomly or otherwise) at the outset of the experiment and appear to have become increasingly unsatisfactory as comparison groups as the experiment progressed.

3.11. The Sister M. Alcuin Study

Chapter 18 in Delacato (1966) is a report of a comparative experiment performed by Sister M. Alcuin of Sacred Heart School in Milwaukee, Wisconsin. This experiment was different from the preceding experiments in that three comparison groups were involved: Experimental Group - 40 students receiving reading instruction and neurological training; Control Group I - 40 students receiving only reading instruction; Control Group II - 40 students receiving reading instruction and some unspecified type of psychological treatment.

Prior to the opening of the six-weeks summer session, the 120 pupils, ranging in age from 6 to 14 years, were tested with the Stanford Reading Achievement Test, the Large-Thorndike Intelligence Test, the Keystone Visual Survey Test and Delacato's tests of laterality and neurological organization.

The manner in which a subject was assigned to either the experimental group or control I or II was not described by the author of the report. It was simply stated that a pupil's score on the "Stanford Reading Achievement Test along with teacher judgment" determined whether he was placed in the experimental or one of the control groups. Prior to the experimental period the following descriptive data were gathered:

	Experimental Group	Control Group I	Control Group II
n	40	40	40
Mean IQ	98	97	97
Mean Age (mos.)	118	125	123

Means for the three groups on the Stanford Reading Achievement Test were not reported, though they must have been available to the author of the report. Although somewhat comparable on IQ, the experimental group was, on the average, seven months younger than Control Group I and five months younger than Control Group II. The amount of discrepancy between the three groups in chronological age reflects on the processes by which subjects were assigned to groups. One might ask for example, "Were the differences in ages about as large as one would expect to observe after random assignment of the 120 subjects to three groups?" A crude and conservative statistical test can be performed which bears on this question. Given that the age range of the 120 subjects was 72 months to 168 months, a reasonable approximation to the standard deviation of the ages would be 13 months. Thus, a reasonable overestimate of the within-groups variance of chronological age would be $13^2 = 169$. For the data on means reported above, the mean square between the three groups for chronological age is 520. The F-ratio $520/169 = 3.1$ can be tested against the 95th percentile in the F distribution with 2 and 117 degrees of freedom; it is significant. Hence, there is less than one chance in twenty that the assignment of the 120 subjects to the treatment groups was like random assignment. (This hypothesis test is meant to be descriptive and was not intended as a test of an hypothesis, viz., that assignment of pupils to groups was random, which was known a priori to be false.) The three comparison groups not only differed by an amount greater than chance expectancy on the chronological age variable, but they could be expected to differ on variables related to chronological age as well.

The average gains on the Stanford Reading Achievement Test over the six-weeks experimental period were as follows: Experimental Group, 0.75 yrs.; Control Group I, 0.40 yrs.; control Group II, 0.44 yrs. An analysis of variance revealed significant differences among the means. Application of the Scheffé method of multiple comparisons showed that the mean gain of the experimental group was significantly different (at the .01 level) from the mean gains of the two control groups. There was no significant difference between the control group means.

The report of this experiment is quite brief and many questions are left unanswered. How did the three groups compare initially on several variables (e.g., reading achievement, motivation) which might be related to the gains they might be expected to make? It was shown above that the Experimental Group was significantly younger than either Control Group. Is this important? How much control was exercised over the instruction in reading during the experiment? Even though students from each group being compared were present in each of six different classrooms, within each class Control Group I was taught by one teacher while the Experimental Group and Control Group II were taught by a different teacher (p. 152.). In what respects and to what extent did the teachers of the Experimental Group and Control Group II differ from the teachers of Control Group I? Such questions do not arise in connection with experiments in which subjects are randomly assigned to experimental conditions and then treated independently with all other influences (teachers, time of day, etc.) held constant or randomized as well. Failure to meet these minimum requirements of a valid comparison casts doubt on the results of an experiment.

3.12. The Miracle Study

The most complete and detailed report of any experiment appears as the last chapter in Neurological Organization and Reading (1966). The experiment was performed and reported by Brian F. Miracle.

Forty students ranging in age from 8 years - 7 months to 11 years - 4 months and reading at least one year below grade level (on the Iowa Test of Basic Skills) were used as subjects. The range of grade-placement reading scores of these fourth and fifth grade pupils was 1.9 yrs. to 4.1 yrs.

Prior to the start of the eight-weeks experimental period, the neurological organization of each pupil was evaluated. Six tests of handedness, four tests of footedness, and five tests of ocular dominance were administered. The number of times each side of the body was employed on all tests was recorded. The absolute value of the difference between the number of times either the left hand, foot or eye was employed and the number of times either the right hand, foot, or eye was employed was taken to be a measure of lateral dominance. High "dominance scores" imply good neurological organization; near zero scores imply poor neurological organization.

Four groups of ten subjects each were compared in the experiment. Group A received reading instruction ("whole" or "sight methods" plus structured and phonetic analysis of words) for thirty minutes daily plus thirty minutes of cross-pattern creeping and cross-pattern walking. The thirty minutes of creeping and walking each day constituted the entire program of neurological training. Group B received only neurological training in the form of thirty minutes of cross-pattern creeping and walking daily; Group B did not receive any reading instruction. Group C received the same remedial reading program as Group A; Group C was not given neurological training. Group D received neither reading instruction nor neurological training. The following diagram summarizes the treatments administered to groups:

	Remedial Reading	No Reading Instruction
Neurological Training	Group A	Group B
No Neurological Training	Group C	Group D

Miracle reported that the group of 40 subjects was divided into four groups of ten students each at random. If so, this experiment represents the only experiment Delacato presented in which random assignment of subjects to groups took place. Hence, Miracle's study would represent the single experiment in which one could be confident that the groups being compared were equivalent (randomly) on all variables at the outset. Unfortunately, either Miracle obtained an unlucky random split of the 40 subjects or else some non-chance factor influenced the assignment of subjects to the four groups. This can be seen from an analysis of the pretest data which Miracle presented in Tables VI, VIII on pages 172 and 176 of Delacato (1966). At the time of the random assignment of the subjects to the four groups, the Iowa Test of Basic Skills (Form 1) was administered. The mean number of items correct for the four groups on the Reading Ability subtest were as follows: Group A = 12.10, Group B = 14.30, Group C = 10.70, Group D = 10.30.

Miracle did not report variances for each group; however he did report the results of "Fisher t-tests" for the pretest data in Table VI. Knowing the means, sample sizes, and value of the *t*-statistic for any comparison of two groups, it is possible to approximate the average within sample variance. This was done for all six comparisons reported among the four groups. The approximation to the average within sample variance turned out to be 14.19. The value of the mean square between the four groups on the pretest of reading ability could be calculated exactly; it proved to be equal to 32.63. The *F*-ratio for testing the hypothesis that the four samples of ten scores each were drawn at random from populations with the same mean is equal to 32.63/14.19 = 2.30. This *F*-ratio exceeds the 90th percentile in the *F*-distribution with 3 and 36 degrees of freedom. Consequently, the differences between the means of the four groups obtained on the pretest were so great that they would occur less than 10 percent of the time, when assignment of subjects to groups is strictly at random. Either the assignment of subjects to groups was not strictly random, as Miracle reported it was, or we must believe that an event occurred when the odds against its occurring were nine-to-one. It should also be noted that the assignment of subjects to groups favored the two experimental groups, Groups A and B. The brighter subjects tended to fall into these groups.

The duration of the experimental period was eight weeks. One cannot learn from the research report whether the four groups had the same or different teachers, whether the groups met at the same or different times of the day, whether the subjects were treated individually or as intact groups. Failure to report such information is a serious omission.*

The *Iowa Test of Basic Skills* (Form 2) was administered at the conclusion of the experimental period. The following data show pretest and posttest means

*Nor did such information appear in Miracle's dissertation, which was also examined.

and mean gain scores on the Reading Ability subtest for each group:

	Group A	Group B	Group C	Group D
n	10	10	10	10
Pretest Mean	12.10	14.30	10.70	10.30
Posttest Mean	18.00	21.10	13.40	12.40
Mean Gain	5.90	6.80	2.70	2.10

Miracle presented *t*-tests of the differences between the posttest means in Table VI. Of course, this multiple *t*-testing of the data is not legitimate and tends to show a spuriously large number of significant differences. An analysis of variance followed by multiple comparisons of the means if the original *F*-test is significant is the appropriate analysis. Though Miracle did not report variances on the posttest of reading ability for the four groups, it was possible to approximate these from a knowledge of means, sample sizes, and *t*-statistics for each pair of groups. The approximation to the mean square within groups obtained from the data in Tables VI and VIII was equal to 24.95. The *F*-ratio for testing the hypothesis of no differences between the four population means was equal to 5.53, which is significant at the .05 level. Tukey multiple comparisons of the posttest means revealed that Groups A and B did not differ significantly from each other but differed significantly from Groups C and D; Groups C and D did not differ significantly.

There is no question that statistically significant differences exist on the posttest of reading ability in the Miracle study. However, legitimate questions remain concerning the initial comparability of the groups (they were significantly different on the pretest of reading ability at the .10 level) and the comparability of the experimental conditions for each group (the four groups may have been treated as intact groups, in which case the experiment did not yield a valid estimate of error). But even apart from these considerations, the data from the experiment were quite surprising. The remedial reading program which was carried out for eight weeks appears to have been ineffectual. Group B made a greater gain (numerically, though the difference is not statistically significant) from pretest to posttest than did Group A, even though Group B received no reading instruction. The average gains shown by Groups C and D, 2.70 points and 2.10 points, were not significantly different, even though Group C was given eight weeks of remedial reading and Group D was given nothing! The average gains for the two experimental groups receiving neurological training in the form of cross-pattern creeping and walking were far greater than the average gains for the two groups which did not receive neurological training. If the experiment was not invalidated by some extraneous influences, we are forced to conclude that 16 hours of cross-pattern creeping and crawling was very effective in improving reading ability and that the remedial reading instruction was quite ineffective. This is indeed surprising. Miracle concluded that cross-pattern creeping and walking alone are more effective in improving reading performance than is remedial reading instruction.

An alternative explanation of the results is not close at hand, though one may exist if it can be determined that the four comparison groups were treated as intact groups studying under different teachers, at different times of the day, etc. Miracle reported as his third conclusion that the "students who showed greatest progress in this study (Groups A and B) were probably more interested in participation than were the students in Group C." This may have been only one of several systematic differences between four intact groups.

4.0. The Correlation of Neurological Organization with Reading Performance and Other Variables

Five empirical studies in Delacato (1959, 1963, 1966) contain information on the correlation of neurological organization and variables such as reading performance, intelligence, and others. It is important to distinguish these correlational or status studies (Stanley, 1961) from the twelve experimental studies reported on in Sections 3.1 - 3.12. While experimental studies in which attempts to manipulate a variable, e.g., by improving the neurological organization of a randomly chosen half of the available subjects, permit some optimism for finding causal links between variables, the status study in which two variables are correlated is less likely to produce valid evidence concerning causality. These considerations are "old-hat" to the research methodologist, but it is often necessary to raise the issue when the opportunity for drawing unwarranted conclusions seems imminent.

One can deduce from Delacato's theory that measures of neurological organization should be closely related to (have a high correlation with) measures of reading performance. It can also be inferred from Delacato's remarks at several points that neurological organization should be more highly related to reading performance than to some other psychological constructs (non-verbal intelligence, for example).

> "Speech and reading are the final human result of neurological organization and hence are clinical indices of the nature and the quality of neurological organization of an individual." (Delacato, 1963, p. 7)

> "Some clinicians have used intelligence tests to draw conclusions. They have analyzed test scores and have made diagnoses of emotional and even organic conditions from this analysis. This naive view of emotional or organic problems is especially prevalent with those who use the Wechsler Intelligence Scale for Children to draw such conclusions." (Delacato, 1966, p. 8)

"Reading is neither a conceptual process nor an intellectual process.... Reading is a perceptual process." (Delacato, 1966, p. 10)

Clearly, the import of these quotations is that neurological organization should be a good indicator of reading performance, but a poorer indicator of general or non-verbal intellectual ability. If reading and speaking are clinical indices of neurological organization and if it is naive to use the WISC to diagnose organic conditions, then we can expect neurological organization to predict reading performance--the "final human result of neurological organization"--better than performance on a non-verbal intelligence test. Such a comparison does not appear explicitly in any of Delacato's three books, perhaps because it would have produced unfavorable results. However, from the data reported in Chapters 10 and 11 of Neurological Organization and Reading, we can test this hypothesis and the primary hypothesis that there should be a substantial correlation between neurological organization and reading performance.

In Chapter 10, Carrick and Watson correlated scores on the adapted Delacato Tests of Neurological Organization and Form J of the Stanford Achievement Test: Elementary Reading Test. For 87 third-grade pupils, r equaled + .35, which is significant at any reasonable level. Aside from the fact that Carrick and Watson were testing a patently true hypothesis to begin with (that neurological organization and reading achievement are correlated), the verification of which could not have great import for the theory, the data indicated clearly a weak, positive relationship.

In Chapter 11, Sister Mariam selected 208 fifth-grade pupils and divided them into Group A "Neurologically Well Organized" (n = 63) and Group B "Neurologically Poorly Organized" (n = 140). The 208 pupils were tested with the Lorge-Thorndike Intelligence Test: Non-Verbal Battery (we shall assume

Level 3, although no indication of which level was used is given in the chapter), the Iowa Tests of Basic Skills, and the Silent Reading Diagnostic Tests. The means and standard deviations for Groups A and B on these tests are given in Table IV (p. 91), Table IX (p. 94) and Table VII (p. 96) as well as in intermediate tables. From these data it is possible to calculate the biserial correlation between measures on the Doman-Delacato Scale for Neurological Organization and measures of non-verbal intelligence and reading achievement.*

In Table IV (p. 91), the following summary statistics are reported:

Group A	Group B
Neurologically Well Organized	Neurologically Poorly Organized
n = 63	n = 140
Mean = 108	Mean = 99
Standard Deviation = 27.14	Standard Deviation = 34.35

First, inspection of the grouped frequency distributions in Table IV (p.91) should reveal to even the most casual reader that the reported standard deviations are greatly in error. The actual standard deviations calculated from the grouped frequency distribution in Table IV are approximately 10.34 for Group A and 10.86 for Group B. (Sheppard's correction for grouping was not employed. Its application would reduce each standard deviation by approximately .10.) It is extraordinary that the reported values could be so greatly in error. The bogus standard deviations in Table IV were used in statistical hypothesis tests on page 91 with misleading results. (One hopes that the other statistics reported

*Biserial *r* estimates the correlation of two variables *X* and *Y* when one is given the measures on *X* and only dichotomous measures (high-low; Group A - Group B) in place of the measures on *Y*.

in Chapter 11 are more reliable.) Using the correct standard deviations, the biserial correlation between neurological organization and non-verbal IQ is .47.*

In Table IX, the mean and standard deviation of scores on the Comprehension scale of the Iowa Test of Basic Skills are 5.3 and 2.57 respectively for Group A and 5.2 and 3.39 for Group B. These values produce a biserial correlation between neurological organization and comprehension (as measured by the Iowa Test of Basic Skills) of .019. The biserial correlation of neurological organization and the Vocabulary Scale of the Iowa Test of Basic Skills was calculated from the same table and proved to be only minutely larger. These essentially zero correlations agree with the non-significant and near zero correlations of neurological organization as manifested in creeping and the reading section of the California Achievement Tests which were reported by Robbins (1966). As the evidence mounts, the correlation of +.35 in Chapter 10 begins to look like the anomaly.

In Table XII the results of testing groups A and B with Bond, Clymer and Hoyt's Silent Reading Diagnostic Tests are reported. The biserial correlations between neurological organization and scores on Comprehension in Isolation and Comprehension in Context are .13 and .15, respectively. The largest biserial correlation between neurological organization and any subtest of the Silent Reading Diagnostic Tests is with Visual Recognition and does not exceed .20.**

*In calculating *r* from the summary statistics it is necessary to employ techniques for finding the variance of combined groups; see Ferguson, 1966, p. 72. Although no exact hypothesis test exists for biserial *r*, a value of .47 for an *n* of 203 is easily significant at the .01 level.

**The critical ratio or z-tests reported in Table XII (p. 96) to test the differences between Groups A and B on the subtests of the Silent Reading Diagnostic Tests are not appropriate since population variances are unknown. Only three of the six tests are significant by the z-test and one of these (Comprehension in Context) becomes nonsignificant at the .05 level ($t = 1.62$ with 201 degrees of freedom) when the appropriate *t*-test is run. (Heterogeneous population variances are a possibility in this example, and because of the unequal sample sizes, the *t*-test is slightly conservative. See Scheffé, 1959, p. 340.)

The conclusion seems obligatory that measures of neurological organization are more highly correlated with measures of non-verbal intelligence than they are with measures of reading achievement. The higher correlation of neurological organization with non-verbal intelligence than with reading achievement cannot be attributed to greater unreliability in the reading tests than in the intelligence tests. The alternate forms reliability for the Lorge-Thorndike Non-Verbal Intelligence Test is .80 for Level 3 at grade 5. The reliability of the Iowa Test of Basic Skills subtests are all around .85 (see p. 86 of Chapter 11); and the corrected split-half reliability of Form J of the Stanford Achievement Test: Elementary Reading is .90; the internal consistency reliabilities of the subtests of the Silent Reading Diagnostic Tests used in Chapter 11 all exceed .70. Although the correlation in Chapter 10 is not strictly comparable with the correlations in Chapter 11 because a different population of subjects was sampled in the two chapters, the correlations within Chapter 11 which showed the greatest discrepancies are comparable.

Delacato's theory definitely does not predict that neurological organization should correlate more highly with non-verbal intelligence than with reading performance. Although the theory is vague on this point, the import of Delacato's writing is that one should expect that non-verbal intelligence would be less highly correlated with neurological organization than would reading achievement. Data which Delacato cited as supporting his theory were shown above to indict the theory. No compelling evidence for a positive correlation between neurological organization and reading performance has been presented. Considerable evidence to the contrary exists. Delacato's theory is neither sufficiently precise in specifying the interrelationships between neurological organization and important psychological constructs nor altogether accurate in the predictions which it can be construed as making. Moreover, the consistency with which zero or very small correlations between neurological organization and reading performance are found reinforces doubts about the validity of the experimental studies in which true gains are maintained to have occurred or in which it is claimed that the superiority of an experimental group over a control group has been established.

Delacato (1963) presented evidence on the relationship between neurological organization and reading performance which must be questioned seriously. Twelve classroom teachers were taught to evaluate neurological organization of pupils in a four-hour orientation session. After the teacher ranked the pupils from highest to lowest on neurological organization, the Stanford Achievement Test was administered to the 248 pupils. A reading performance score was found by averaging the "reading comprehension, vocabulary, and spelling scores" of the pupils. (It will be assumed that grade-placement scores are the "scores" to which Delacato refers, since it appears to be his habit to measure reading performance in this manner.) Delacato reported five correlations of "teachers' evaluation of neurological organization" and "reading performance": .72, .87, .81, .64, and .84. I have found it impossible to ascertain from Delacato's report why there are five different correlations. In the first sentence of the third paragraph on page 138, a single reference to "the sections were placed in order from the most organized child...to the least organized in each section." This is the only reference to "sections" in the report. I have no idea what the sections are, how they were formed, or how many of them there were. It may be that the five separate correlations arose from five "sections."

Are there important implications from these high correlations of "teachers' evaluation of neurological organization" and "reading performance"? Do they not contradict the small and near zero correlations that have been gleaned from the studies in Neurological Organization and Reading and which appear in Robbins (1966)? Do they not imply that poor neurological organization produces poor reading? The answer to all three questions is "Probably, no." In the first place, correlation cannot be taken as direct evidence for causation; but a lack of correlation begins to raise doubts about claims for a causal link between the two variables. Secondly, to the extent that teachers' evaluations were subjective and to the extent that they knew their evaluations were being gathered for the purpose of studying reading problems, spurious positive correlations could arise. Moreover, these conditions could be expected to exist in some appreciable degree, if the teachers were not fully unaware that Dr. Delacato believed that there should be a high relationship between neurological organization and reading performance. However, this point will not be belabored since the spuriousness of the correlations appears to stem from a more obvious source. Delacato's correlations were not calculated on a homogeneous sample of pupils. Chronological ages ranged from six to thirteen years; IQ's ranged from 96 to 149. Obviously, one would expect to find large positive correlations between two variables which are both a function of age--even if the variables are unrelated at any given age--if the age variable is allowed to vary. Older children have better coordination than younger children; older children read at higher grade levels than younger children. If neurological organization and reading performance are correlated over a large span of chronological ages, a large positive correlation will result. The same is true when height and mental age are correlated across different chronological ages because each is a function of aging. If Delacato's correlations are to have any implications for the study of how reading performance might be affected by neurological organization, they must be calculated at each chronological age separately or better, the correlation must have chronological age partialed out of it. When this is done, it is likely that those high correlations of around .80 will fall in line with the other near zero correlations of neurological organization and reading performance which have been found repeatedly when correlations are calculated on pupils of homogeneous chronological ages.

Miracle reported correlations between neurological organization and reading ability in Chapter 19 of Neurological Organization and Reading. As with Delacato's study of the correlations of the two variables, the chronological ages of the subjects were allowed to vary. As nearly as one can tell from Miracle's report, the ages of the 40 pupils on which his correlations are based, ranged from 8 years 7 months to 11 years 4 months. Chronological age was not partialed out of any of Miracle's correlations; hence we would expect them to be spurious because of an obvious common relationship of reading performance and neurological organization to age. Half of Miracle's 40 subjects received Delacato therapy; the other half were used as controls. Reading performance scores on the Iowa Basic Skills Test given at the end of the experiment correlated .168 and .242 with composite dominance scores taken before and after the experiment, respectively* These correlations are not statistically significant, a value of .304 being required for significance at the .05 level with a two-tailed test. (Miracle's reported critical value of .315 for significance at the .05 level is slightly in error.) Miracle made an astounding interpretation of this failure to find correlation between laterality and reading performance.

> "That the posttest reading scores showed no significant relationship with either the pre or posttest composite scores of dominance was not unexpected in this study. The purpose of this study and the manner in which it was conducted allowed for only fifty percent of the subjects, or twenty of the forty students, to strengthen in any way their dominance. In light of this, these findings seem consistent with expectations." (Delacato, 1966, p. 170)

*The stability reliability of the "composite dominance" score was .943.

Miracle's interpretation and expectations are incorrect. If neurological organization and reading performance are in fact related to each other, increasing the neurological organization of only a portion of the subjects while leaving the remainder untreated should <u>increase</u> the correlation. Most elementary statistics students would recognize this as increasing a correlation by creating identifiable subgroups, viz., neurologically <u>well</u> and neurologically <u>poorly</u> organized, within a sample. (The reader is referred to pages 166-167 of Walker and Lev, <u>Elementary Statistical Methods</u>, 1958.)

Kabot (Delacato, 1966, Chapter 14) presented data from which one can reconstruct information on the correlation of reading performance with neurological organization. From 167 third-grade pupils, Kabot selected 43 who were "over-achievers" and 46 who were "under-achievers," as defined by the relationship between <u>Kuhlmann-Anderson IQ Test</u> scores and <u>Stanford Reading Achievement Test</u> scores. Approximately 27 percent of the pupils were classified as under-achievers and 27 percent as over-achievers. The following contingency tables can be constructed from the first paragraph on page 119 of Delacato (1966):

Table 1

Neurological Organization (Defined by Mixed Hand and Eye Dominance)	Reading Performance: Underachievement	Reading Performance: Overachievement	
Good	18	31	49
Poor	28	12	40
	46	43	89

Table 2

Neurological Organization (Defined by Mixed Hand-Eye-Foot Dominance)	Reading Performance: Underachievement	Reading Performance: Overachievement	
Good	7	17	24
Poor	39	26	65
	46	43	89

Fortunately, Kabot chose quite by accident to look at the upper and lower 27 percent of the pupils on the reading performance dimension. Consequently, we can gain information about the correlation of neurological organization and reading performance by referring the percents of persons occupying certain cells in Tables 1 and 2 to well-known tables for finding _r_ from the "upper and lower 27 percent-groups." (See for example, Clover, 1959.)

The data in Tables 1 and 2 lead to the following estimates of the correlation between continuous measures of neurological organization and reading "overachievement" and "underachievement": for Table 1, _r_ equals .34; for Table 2, _r_ equals .30. An asymptotic approximation to the standard error of _r_ found in this manner has been determined by Ross and Weitzman (1964). If the population correlation is actually zero, _r_ determined by the 27 percent method will have a standard error of approximately .105 for an _n_ of 167. Thus the two _r_'s above are statistically significant.

In Table 3 is reported the available information about correlations between neurological organization and reading performance and the manner in which the sizes of these correlations are related to the chronological age range of the sample of subjects. There is a clear indication in Table 3 that the size of r is a function of the heterogeneity of chronological ages in the sample. Of the 24 values of r, 15 values--all of them computed on samples with an age range of one year--fall into the limbo of scientific insignificance, not to mention statistical insignificance.

The logic of partialing chronological age out of the correlation between neurological organization and reading performance should be subjected to scrutiny. Two situations can be identified which lead to different recommendations regarding the use of partial correlation in this situation:

1. If the aspects of neurological organization which are relevant to reading performance mature with age and this maturational process cannot be speeded by therapy, then it is logical to partial chronological age out of the correlation between the two variables.
2. If the aspects of neurological organization which are relevant to reading performance mature which age and this maturational process can be speeded by therapy, then it is not defensible to partial chronological age out of the correlation.

We have seen that when chronological age is partialed out of correlations of neurological organization and reading performance, they drop--if they are not already essentially zero--to values which are scientifically insignificant. For a given chronological age, neurological organization accounts for only about five percent of the variance, most of which is reliable, in reading performance. However, when chronological age is allowed to vary, the correlation between neurological organization and reading performance increases. It is illuminating

Table 3

Relationship Between the Correlation of Neurological Organization with Reading Performance and the Range in Chronological Age in the Sample.

Study	Range of Chronological Ages in Sample	n	Number of r's in the study	Values of r obtained
1. Robbins (1966)	1 year	126	4	-0.01 to 0.05
2. Sister Mariam (Chapter 11 in Delacato 1966)	1 year	203	8	0.02 to 0.15
3. Kabot (Delacato, 1966, Chapter 14)	1 year	167	2	0.30 to 0.34
4. Carrick and Watson (Chapter 10 in Delacato 1966)	1 year	87	1	.35
5. Miracle (Chapter 19, Delacato 1966)	2 yrs. 9 mos.	40	4	.17 to .47
6. Delacato (1963, pp. 136-138)	7 years	248	5	.64 to .87

to cite Miracle's study again at this point (Delacato, 1966, Chapter 19). Prior to an experiment in which 20 of 40 subjects (with an age range of 2 years 9 months) received Delacato therapy, the correlation of "reading ability and composite scores of dominance" was .442. If Delacato therapy increases neurological organization and this increase has a concommitant effect on reading performance, the correlation of the two variables should be higher after the experiment than before. (In other words, the correlation should be like one calculated on subjects with a greater age range since Delacato therapy is supposed to increase the neurological organization of half of the subjects just as maturation over time does.) In fact, the correlation between "reading ability and composite scores of dominance" is only .242 after the experiment. We have seen that the reasoning Miracle presents to explain this discrepancy is fallacious.

The significant positive correlations which Delacato presented between neurological organization and reading performance appear to be artifactual. When properly regarded, they are consistent with the non-significant correlations found between reading performance and aspects of neurological organization by Balow (1963), Balow and Balow (1964), Flescher (1963), Coleman and Deutsch (1964), and Silver and Hagin (1960), and Hillerich (1964).

Balow (1963) found no significant correlation in a group of 302 first-grade pupils between reading achievement and any of the following variables: hand dominance, eye dominance, hand-eye dominance, and strength of dominance. In 1964, Balow and Balow observed 250 second-grade pupils and found no significant differences in reading achievement between a group of 140 pupils having hand and eye preference on the same side of the body, a group of 87 pupils having hand and eye preference on opposite sides of the body, and a group of 23 pupils having mixed hand preference. From Table 1, the correlation ratio, or eta-squared, can be reconstructed from the reported means and F-ratios. The value of the correlation ratio for the variable "word reading" is +0.015; the value of the correlation ratio for "paragraph reading" is even closer to zero. (For an interpretation of the correlation ratio, see McNemar, 1955, Chapter 15.)

Hillerich followed 400 pupils from Kindergarten through the second grade. When classified into the four groups (a) right (R) eye - right (R) hand dominance, (b) R eye - left (L) hand dominance, (c) L eye - R hand dominance and (d) L eye - L hand dominance, only small numerical and nonsignificant differences where observed between the groups on reading achievement, reading differential, IQ, and reading reversal variables. A group of 57 pupils who did not show an unequivocal dominance of either hand or either eye had the lowest mean scores on all four dependent variables, but in no instance did the mean of this group differ significantly from the mean of the right eye - right hand dominance group. It must also be noted that in evaluating Hillerich's results that two-thirds of the pupils below grade level in reading achievement were boys. A careful correlational analysis of Hillerich's data would entail calculating correlations for boys and girls separately. Distinguishing the sexes would lower all correlations.

Capobianco (1966) studied 38 males and 20 females ranging in age from 153 months to 200 months. The group of 58 pupils was divided into 34 (26 males and 12 females) who had established laterality patterns and 24 (13 males and 11 females) who had not. Scores on the reading subtest of the Wide Range Achievement Test showed that the non-established group actually scored <u>higher</u> (though the difference was statistically nonsignificant) than the group in which laterality patterns were established. The superiority (numerical but not statistically significant) of the non-established group held up when the data were analyzed separately for males.

That Delacato and the authors of the studies which he reprinted did not refer to any of the published studies which show results contradicting their own was a major oversight.

It should be clear that the correlational approach, whether supplemented by partialing out other variables or not, cannot give a definitive answer to the question whether or not Delacato's therapy improves neurological organization in ways relevant to reading performance. Only <u>experiments</u> in which attempts are made to improve reading by improving neurological organization can yield a trustworthy answer. None of the experiments which Delacato esteems for "excellence of design and control" is, in fact, sufficiently well designed, controlled, analyzed, or reported that it constitutes a valid piece of scientific evidence.

5.0. Conclusion

It has been the spirit of this review to attempt to get at the truth. I have no interest vested in any particular method of reading instruction. My only interest is in cutting through to the truth which seems to lie beneath empirical, experimental data. In the evaluation of each study reported in Delacato's three books, an attempt was made to avoid caviling at and carping about the myriad statistical faux pas which were of no consequence to getting the "sense" out of the data. Nor was the game of conjuring up improbable alternative explanations of gains and differences played. Whenever an attempt was made in the review of a study in the paper to explain the obtained results in terms of factors other than the effectiveness of Delacato's therapy, the explanation offered was considered probable and not simply in the realm of possibilities. In many instances, possible explanations of gains shown by a treated group or of a difference favoring the experimental group over the control group were rejected and not mentioned because they seemed to be only possibilities.

To argue that "possibilities" always exist which could render any experiment invalid by my standards would be to misinterpret the position taken in this review. The fact is that a valid, controlled, and crucial experiment to evaluate the effectiveness of Delacato's system of therapy can be executed at reasonable expense of time and money. There is no need to outline the design of such an experiment here. It could be designed by anyone who understands and appreciate the necessity for random assignment of experimental units to the experimental and control groups, for treatment of the experimental units--

whether they be classrooms or individuals--independently and identically (or in ways which differ only randomly) except for the presence of the therapy in the experimental group, and for an appropriate statistical analysis of the data.

Serious doubts have been raised in this review about the results of practically every empirical study which Delacato has cited as support for his theory of neurological organization and reading. In my opinion all but three or four of the studies are beyond redemption. These last experiments were so poorly designed and executed--so many extraneous influences were confounded with the experimental variable--that nothing short of new, more adequately designed experiments can resolve the question of the validity of their conclusions.

Without an exception, the empirical research in Delacato (1959, 1963, 1966) in poorly reported. It cannot be argued that the omission of such data was necessary to reduce the length of the reports, because the same research reports are littered with irrelevant observations, homolies, and panegyrics on Delacato therapy. Either the authors of these reports had no "feel" for which data are important indications of the validity of their experiments or, for one reason or another, they chose not to report such data. For example, in Chapter 18 of Delacato (1966), the experimental and two control groups were matched at the beginning of the experiment on age and Lorge-Thorndike IQ. Even though the Stanford Reading Achievement Test was given both before and after the experimental period and was used to assess "gains," the pretest means on the reading achievement test for the three groups are not reported. Only the average gains for each group are given. Obviously the means for each group on the Stanford Reading Achievement Test could have been reported with little effort. Why was this not done? In Chapter 14 of Delacato (1966), a significant difference was shown between the experimental and control group only after a second posttest which followed the experiment by one year and after four of the original eleven matched pairs had been dropped from the study. No explanation as to why four of the matched pairs were dropped was given. In the four matched pairs that were dropped--regardless of why they were missing in the final analysis--did the control subject tend to make a greater gain than the matched experimental subject on the first pretest? In the Glaeser study (Chapter 16 in Delacato, 1966) no explanation of the incomplete data was given even though as many as eight experimental and two control subjects were unaccountably missing.

No position has been taken in this review on the question of the viability of Delacato's theory of neurological organization nor the question whether Delacato's therapy for improving reading performance will ultimately prove to be effective in an improved form or with special children. These are substantive questions as opposed to questions of the methodology and techniques of empirical research. The position taken here is that extravagant claims have been made for the validity of experiments which Delacato has reported as supporting his claims. Without exception, these experiments contained major faults in design and analysis. About half of the experiments were so inadequate that they are not acceptable as evidence by the standards against which educational research is presently evaluated. Sources of bias and probable invalidity have been identified in the remaining experiments which make the reported results questionable. At best, uncontrolled factors inflated small, but legitimate effects due to Delacato's therapy in each of the experiments; at worst, these uncontrolled influences were the sole sources of gains or differences between experimental groups. Either extreme is possible. Enough doubt has been cast on the results of all of the experiments that either replications of them under improved conditions or the publication of adequate research reports will be required before the conclusions drawn from them are admissible. Then the debate can proceed on the basis of sound empirical research, as it must. Then it will be appropriate to open the discussion to other research in laterality and to experimental tests of Delacato's therapy not acknowledged by him which have shown negative results (see, for example, Robbins 1966) under conditions of control equal to and surpassing those which prevailed in the experiments reviewed in this paper.

If clarifications and more complete research reports are not forthcoming from Dr. Delacato and the authors of the studies published in his books, steps should be taken by individual reading researchers to examine the original data and report their findings. The case of Dr. Berardine Schmidt (1946) ought to be fresh in the memory of all educational researchers. Dr. Schmidt purported to have brought about incredible changes in the personal, social, and intellectual behavior of "feebleminded" children. Through the efforts of Samuel A. Kirk (1948) the results Dr. Schmidt claimed to have obtained were discounted, and now appear to have been quite fallacious, because of Dr. Schmidt's unorthodox research techniques and reporting which bordered on unethical practice. It is an interesting phenomenon which can only be ascribed to "wishful thinking" and the effectiveness of mass media that Schmidt's thoroughly discredited findings are still being cited as the product of valid research and that Kirk's exposé is seldom noted (see Thomasson and Stanley, 1955).

Not being thoroughly acquainted with Delacato's position, the reader might have assumed falsely that therapy to bring about neurological organization should show effects only on persons suffering serious neurologically dysorganization initially. It would follow from this assumption that any study in which Delacato therapy was given to normal pupils would constitute neither a valid test of his theory nor an appropriate evaluation of Delacato's therapy. However, with remarkable sanguinity, Delacato has argued that this therapy should be effective for both the neurologically dysorganized (including the brain injured, genetically deficient, and environmentally deprived) and the pupil whose neurological organization is normal.

"The author further feels that a child with good reading can be helped to have even better language facility and better language aptitudes through the system of setting up a neurological organization which operates as a unity. No doubt as man has evolved he has set up certain environmental blocks to his complete utilization of his neurological structure. Hence if we can, through preventative activity or through educative activity, teach people neurological unity, we shall have done them a great service and shall perhaps make our good students even better, our good language people (sic) even better, our good spellers even better, our fluent speakers and listeners even better. Indeed we may be discussing a means for hurrying the evolutionary process." (Delacato, 1959, p. 80)

"...the author feels that the approach used above and the results thereof certainly indicate that the rationale contained herein is quite applicable to the normal classroom activity for children who present slight deviations in reading as well as for children who present gross reading retardation." (Delacato, 1959, p. 100)

It seems advisable that at least two distinct groups of subjects be identified in any experiment on the effect of neurological organization on reading performance: "normal" pupils who do not give evidence of marked neurological dysorganization, and pupils who possess marked neurological dysorganization. It is important to distinguish these two subgroups because the effects of the Delacato therapy may not be the same for each group. Delacato has maintained that his therapy is effective on both groups (see <u>The Treatment and Prevention of Reading Problems</u>, p. 80). It may be, however, that only the markedly neurologically dysorganized can benefit from attempts to establish hemispheric dominance and other conditions which constitute adequate organization. More adequate designs than those which have been employed thus far in investigating the effects of Delacato's therapy would involve stratifying the sample of subjects into at least the two groups mentioned at the beginning of this paragraph and looking for the possible differential effectiveness of the therapy. This has not been done in any of the studies reviewed in this paper. The subjects who participated in almost all of the experiments reviewed in this paper could not be characterized as seriously neurologically dysorganized. A generous assessment of the research Delacato cites as evidence for the effectiveness of his therapy might be as follows: all of the empirical research reported thus far has failed to produce cogent evidence that Delacato's therapy has any effect whatsoever on the reading performance of <u>normal subjects</u>; the possibility exists that Delacato's therapy is effective on subjects suffering serious neurological dysorganization, though this hypothesis has not been subjected to adequate empirical tests. If it were to be reliably and validly established that the highly neurologically dysorganized child could be rehabilitated as Delacato maintains, it would represent a truly valuable contribution to the techniques of remediation of certain special learning difficulties. Of course, we would have to relinquish hope that a "means for hurrying the evolutionary process" had been found, or that Delacato's neurological exercises can make good readers even better; but then we should all be accustomed to having our Utopian dreams dispelled by the intransigent facts of life by now.

References

Balow, I. H. Lateral dominance characteristics and reading achievement in the first grade. Journal of Psychology, 1963, 55, 323-328.

Balow, I. H. and Balow, B. Lateral dominance and reading achievement in the second grade. American Educational Research Journal, 1964, 1, 139-143.

Biaggio, L. I. and Stanley, J. C. Regression toward the mean due to errors of measurement. Paper presented at the 9th Inter-American Congress of Psychology, Miami, 1964. (16 pages).

Campbell, D. T. Factors relevant to the validity of experiments in social settings. Psychological Bulletin, 1957, 54, 297-312.

Campbell, D. T. and Stanley, J. C. Experimental and quasi-experimental designs in research on teaching, Chapter 5 in Gage, N. L. (Ed.) Handbook of Research on Teaching, Chicago: Rand McNally, 1963.

Capobianco, R. J. Ocular-manual laterality and reading in adolescent mental retardates. American Journal of Mental Deficiency, 1966, 70, 781-785.

Clover, R. M. Estimating item indices by nomographs. Psychometrika, 1959, 24, 179-185.

Delacato, C. H. The Treatment and Prevention of Reading Problems. Springfield, Illinois: Charles C. Thomas, 1959.

Delacato, C. H. The Diagnosis and Treatment of Speech and Reading Problems. Springfield, Illinois: Charles C. Thomas, 1963.

Delacato, C. H. Neurological Organization and Reading. Springfield, Illinois: Thomas Books, 1966.

Fisher, R. A. The Design of Experiments. (1st Ed.) London: Oliver & Boyd, 1935.

Flescher, I. Ocular-manual laterality and perceptual rotation of literal symbols. Genetic Psychology Monographs, 1962, 66, 3-48.

Hillerich, R. L. Eye-hand dominance and reading achievement. American Educational Research Journal, 1964, 1, 121-126.

Kirk, S. A. An evaluation of the study by Bernardine G. Schmidt entitled: "Changes in personal, social, and intellectual behavior of children originally classified as feebleminded." Psychological Bulletin, 1948, 45, 321-333.

Lindquist, E. F. Design and Analysis of Experiments in Psychology and Education. Boston: Houghton Mifflin, 1953.

Lord, F. M. The measurement of growth. Educational Psychological Measurement, 1956, 16, 421-437.

Lord, F. M. Further problems in the measurement of growth. Educational Psychological Measurement, 1958, 18, 437-451.

Lord, F. M. Elementary Models for Measuring Change, Chapter 2 in C. W. Harris (Ed.) Problems in Measuring Change. Madison: University of Wisconsin Press, 1963.

McDonald, A. S. The Placebo Response in Reading Research. Pages 220-229 in New Developments in Programs and Procedures for College--Adult Reading, 12th Yearbook of the National Reading Conference, 1963.

McNemar, Q. Psychological Statistics (2nd Ed.) New York: John Wiley, 1955.

McNemar, Q. On growth measurement. Educational and Psychological Measurement, 1958, 18, 47-55.

Page, E. B. "Recapturing the richness within the classroom." Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, Illinois. February 11, 1965. (7 pages)

Robbins, M. P. The Delacato interpretation of neurological organization: an empirical study. Unpublished doctoral dissertation, University of Chicago, 1965.

Robbins, M. P. A study of the validity of Delacato's theory of neurological organization. Exceptional Children, April 1966, pp. 517-523.

Ross, J. and Weitzman, R. A. The twenty-seven percent rule. Annals of Mathematical Statistics, 1964, 35, 214-221.

Rulon, P. J. Problems of regression. Harvard Educational Review, 1941, 11, 213-223.

Scheffe, H. The Analysis of Variance. New York: John Wiley & Sons, 1959.

Schmidt, Bernardine G. Changes in personal, social, and intellectual behavior of children originally classified as feebleminded. Psychological Monograph, 1946, 60, No. 5, 1-144.

Stanley, J. C. Studying status versus manipulating variables. In R. O. Collier and S. M. Elam (Eds.), Research Design and Analysis: The Second Phi Delta Kappa Symposium on Educational Research. Bloomington, Indiana: Phi Delta Kappa, 1961, pp. 173-208.

Stanley, J. C. and Beeman, Ellen. Restricted Generalization, Bias, and Loss of Power, That May Result from Matching Groups. Psychological Newsletter, New York University, 1958, 9, 88-102.

Thomasson, Peggy and Stanley, J. C. Uncritical Citation of Criticized Data. Science, April 22, 1955, Vol. 121, No. 3147, pages 610-611.

Thorndike, R. L. Regression fallacies in the matched groups experiment. Psychometrika, 1942, 7, 85-102.

Thorndike, R. L. The Concepts of Over and Underachievement. New York: Columbia University, 1963.

Walker, Helen M. and Lev, J. Elementary Statistical Methods (Revised edition). New York: Holt and Co., 1958.

Wold, H. Causal inference from observational data, Journal of the Royal Statistical Society (Series A), 1956, 119, Part 1, 28-50.