DOCUMENT RESUME

ED 046 892	24	SP 004 640

AUTHOR	Elashoff, Janet Dixon; Snow, Richard F.
TITLE	A Case Study in Statistical Inference: Reconsideration of the Rosenthal-Jacobson Data on Teacher Fxpectancy.
INSTITUTION	Stanford Univ., Calif. Stanford Center for Research and Development in Teaching.
SPONS AGENCY	Office of Education (DHFW), Washington, D.C. Bureau of Research.
REPORT NO	TR-15
BUREAU NO	BR-5-0252
PUB DATE	Dec 70
CONTRACT	OEC-6-10-078
NOTE	173p.
EDRS PRICE	EDRS Price MF-$0.65 HC-$6.58
DESCRIPTORS	*Academic Achievement, *Fducational Research, Elementary School Students, Elementary School Teachers, *Expectation, Prediction, Research Design, Research Methodology, *Research Problems, Statistical Analysis, *Teacher Attitudes
IDENTIFIERS	Pygmalion in the Classroom

ABSTRACT

This paper presents a critical evaluation of the research study Pygmalion in the Classroom by R. Rosenthal and L. Jacobson (New York: Holt, Rinehard and Winston, 1968) and reports an extensive reanalysis of the Rosenthal-Jacobson data. The Pygmalion study purported to show that children whose teachers expected them to "bloom" intellectually would do so. The critique suggests that the Rosenthal-Jacobson report as a whole is inadequate. Descriptions of design, basic data, and analysis are incomplete. Inconsistencies between text and tables, overly dramatic conclusions, oversimplified, inaccurate, or incorrect statistical discrssions and analyses all contribute to a generally misleading impression of the study's results. In their reanalyses of the Rosenthal-Jacobson data, the present authors demonstrate a wide variation in apparent results which can be obtained from slightly different statistical approaches if serious imbalance in design and major measurement problems exist in a research study. They conclude that the reanalysis reveals no treatment effect of "expectancy advantage" in grades 3 through 6. The first and second graders may or may not exhibit some expectancy effect, but a conclusive analysis of first- and second-grade IQ scores is not possible. (Author/RT)

Technical Report No. 15

A CASE STUDY IN STATISTICAL INFERENCE: RECONSIDERATION OF THE ROSENTHAL-JACOBSON DATA ON TEACHER EXPECTANCY

Janet Dixon Elashoff and Richard E. Snow

School of Education
Stanford University
Stanford, California

December 1970

Published by the Stanford Center for Research and Development in Teaching, supported in part as a research and development center by funds from the United States Office of Education, Department of Health, Education, and Welfare. The opinions expressed in this publication do not necessarily reflect the position, policy, or endorsement of the Office of Education. Earlier work on this study was supported by Contract No. OEC4-6-061269-1217 from the Cooperative Research Branch of the U. S. Office of Education. Final work on the study was supported by Contract No. OE-6-10-078, Project Nos. 5-0252-0501, 0704.

This limited edition of 250 copies is being distributed without charge to selected persons. It is the intent of the Stanford Center for Research and Development in Teaching to explore the possibility of subsequent publication of the work under copyright in accordance with the policies set forth in the U. S. Office of Education Copyright Guidelines effective June 8, 1970.

Introductory Statement

The Center is concerned with the shortcomings of teaching in American schools: the ineffectiveness of many American teachers in promoting achievement of higher cognitive objectives, in engaging their students in the tasks of school learning, and, especially, in serving the needs of students from low-income areas. Of equal concern is the inadequacy of American schools as environments fostering the teachers' own motivations, skills, and professionalism.

The Center employs the resources of the behavioral sciences--theoretical and methodological--in seeking and applying knowledge basic to achievement of its objectives. Analysis of the Center's problem area has resulted in three programs: Heuristic Teaching, Teaching Students from Low-Income Areas, and the Environment for Teaching. Drawing primarily upon psychology and sociology, and also upon economics, political science, and anthropology, the Center has formulated integrated programs of research, development, demonstration, and dissemination in these three areas. In the Heuristic Teaching area, the strategy is to develop a model teacher training system integrating components that dependably enhance teaching skill. In the program on Teaching Students from Low-Income Areas, the strategy is to develop materials and procedures for engaging and motivating such students and their teachers. In the program on Environment for Teaching, the strategy is to develop patterns of school organization and teacher evaluation that will help teachers function more professionally, at higher levels of morale and commitment.

This report is a critique and reanalysis of the study of teacher expectancy reported in Pygmalion in the Classroom by Robert Rosenthal and Lenore Jacobson. The importance of the present work derives from the proposition that understanding the role of teacher expectancy in American schools is central to the improvement of teaching.

Acknowledgments

Drs. Rosenthal and Jacobson have cooperated in providing copies of their original data and permission to reanalyze them. We gratefully acknowledge their assistance. Portions of Pygmalion in the Classroom: Teacher Expectation and Pupils' Intellectual Development, by Robert Rosenthal and Lenore Jacobson. Copyright (c) 1968 by Holt, Rinehart and Winston, Inc. were reprinted by permission of Holt, Rinehart and Winston, Inc.

Portions of the work described in this report were supported within a USOE sponsored project on the nature of aptitude (OEC 4-6-061269-1217), and portions were supported by the Stanford School of Education and the Stanford Center for Research and Development in Teaching.

The authors wish to thank N. L. Gage, L. J. Cronbach, and Ingram Olkin for their helpful comments and criticisms during various stages of the work. The assistance of Bruce Bergland, John Burke, James Eusebio, Catherine Liu, Akimichi Omura, Donald Peters, and Trevor Whitford is gratefully acknowledged. Many others have offered helpful suggestions on the manuscript.

Table of Contents

Page

Table of Contents (Continued)

Table of Contents (Continued)

List of Tables

List of Tables (Continued)

List of Figures

Preface

Increasingly, investigators are attempting research on difficult human problems. Many students in education and the behavioral sciences are preparing for research careers. Others are being called upon to read and use the results of research in practice. To these ends, textbooks and courses on research methodology abound. Some aim only at introductions to measurement, experimental design, and statistical analysis. Others prepare the investigator for planning, conducting, and reporting his own research. But textbook examples usually show only orderly and correct results. Seldom is the student confronted with the difficult problems of conducting, analyzing, or criticizing real research data. Discussions of alternative methods and bases for distinguishing among possibly appropriate procedures are usually sketchy and not accompanied by detailed examples. Direct attempts at developing critical and evaluative skills are rare.

This report, a case history of a data analysis, is intended to serve as a special kind of supplement to courses on research methodology and statistical analysis, for the student and the practicing researcher or educator. It is a detailed criticism and case history of a data analysis. At one level, it is a critical evaluation of a research report. At another level, it is a detailed account of technical issues important in evaluating research. At still another, it is a comparison of the merits of, and the results obtained from, alternate analytic approaches to the same data.

The report is a case study of the research study Pygmalion in the Classroom by Rosenthal and Jacobson (1968) and the report of an extensive reanalysis of the Rosenthal and Jacobson data. This study was chosen for detailed examination for two reasons. First, it addresses a major social problem, has received nationwide attention, and has prompted a number of similar studies in the area. Second, its basic design, measurement problems, and the statistical procedures used in its analysis and reanalysis are typical of those encountered frequently in educational or behavioral science research.

J. D. Elashoff

R. E. Snow

Abstract

This report is a critical evaluation of the research study Pygmalion in the Classroom by Rosenthal and Jacobson (1968) and the report of an extensive reanalysis of the Rosenthal and Jacobson data. The Rosenthal and Jacobson study was chosen for detailed examination for two reasons. First, it addresses a major social problem, has received nationwide attention, and has prompted a number of similar studies in the area. Second, its basic design, measurement problems, and the statistical procedures used in its analysis and reanalysis are typical of those encountered frequently in educational or behavioral science research.

Our criticism and reanalysis is intended to serve several purposes. Its major aim is to provide a pedagogical aid for students, researchers, and users of research. Thus it offers an extensive critique of a study, its design, analysis, and reporting. This critique provides a vehicle for examining common methodological problems in educational and behavioral science research, and for discussing and comparing statistical methods which are widely used but seldom well understood. The reanalysis of the Rosenthal-Jacobson data provides a demonstration of the wide variation in apparent results possible when similar analytic procedures are applied to data with sampling and measurement problems. Finally, we sought to identify the conclusions that can reasonably be drawn about teacher expectancy from the Rosenthal-Jacobson study, since the wide publicity attracted by the study's expectancy hypothesis may have already sensitized teachers to this type of experiment and thus prejudiced attempts at replication.

A CASE STUDY IN STATISTICAL INFERENCE: RECONSIDERATION OF THE ROSENTHAL-JACOBSON DATA ON TEACHER EXPECTANCY

Janet Dixon Elashoff and Richard E. Snow

CHAPTER I: INTRODUCTION

This report is a critical evaluation of the research study reported by Rosenthal and Jacobson (1968b) and the report of an extensive reanalysis of their data.

In his 1966 book, Robert Rosenthal, a Harvard social psychologist, demonstrated the importance of experimenter effects in behavioral research, thereby developing a new field for psychological inquiry (Rosenthal, 1966). After a discussion of the experimenter as biased observer and interpreter of data, and of the effects of relatively permanent experimenter attributes on subjects' responses, a series of experiments was summarized purportedly showing the effects of experimenter expectancy in studies of both human and animal behavior. Many suggestions were offered on the control and reduction of self-fulfilling prophecies in psychological research. To suggest the generality and importance of such phenomena, the book closed with a preliminary analysis of data on teacher expectancy effects and pupil IQ gains in elementary school. Those closing pages (pp. 410-413) then were expanded by Rosenthal and Jacobson for journal presentation (1966, 1968a) and for wider circulation in book form (1968b). For brevity in the present report, we will refer to the original study, authors, and book source _Pygmalion in the Classroom_ as RJ.

Our criticism and reanalysis is intended to serve several purposes. Its major aim is to provide a pedagogical aid for students, researchers, and users of research. Thus it offers an extensive critique of a study, its design, analysis, and reporting. This critique provides a vehicle for examining common methodological problems in educational and behavioral science research, and for discussing and comparing statistical methods which are widely used but seldom well understood. The reanalysis of the RJ data provides a demonstration of the wide variation in apparent results when similar analytic procedures are applied to data with sampling and measurement problems. Finally, we sought to identify the conclusions that can reasonably be drawn about teacher expectancy from the RJ study, since the wide publicity attracted by the study's expectancy hypothesis may have already sensitized teachers to this type of experiment and thus prejudiced attempts at replication.

For pedagogical purposes, we have included criticisms ranging from major to relatively minor issues, from points of general information readily available to most educational researchers, to points buried in the statistics literature. It might be argued that our criticisms are unnecessarily stringent, that faults in the RJ study are common faults or that RJ use procedures consistent with "standard practice" in the field. Even if one feels that RJ should not themselves be unduly criticized for faults common in standard practice, one must begin somewhere to examine and improve standard practice. We can see no better place to begin than with a widely quoted popular book that is also "... intended for students of education and of the behavioral sciences, generally, and for research investigators in these fields" (RJ, p. viii).[†]

[†] From Pygmalion in the Classroom: Teacher Expectation and Pupils' Intellectual Development, by Robert Rosenthal and Lenore Jacobson. Copyright (c) 1968 by Holt, Rinehart and Winston, Inc. Reprinted by permission of Holt, Rinehart and Winston, Inc. This credit line applies to all quotations from this source identified in the text by the initials RJ, a page reference, and the symbol (†).

Our report is organized as follows. In the remainder of Chapter I we summarize the RJ study, data analysis, and conclusions. Next, we provide a brief preview of the contents of later chapters. In Chapter II, criticisms of the RJ book as a report of research are discussed. In Chapter III, we discuss design and sampling problems inherent in the RJ study. The fourth chapter deals in detail with the measurement problems encountered in the study. Chapter V examine's RJ's statistical analysis, discusses the difficulties associated with choosing appropriate analytic techniques for such data and presents the main details of our reanalyses. Selected information from the reanalysis is also included elsewhere throughout the report, wherever pertinent. Finally, we review the conclusions that seem warranted by the RJ study and present some methodological recommendations. Brief descriptions of the statistical techniques discussed in the book are included in the appendix.

Summary of the RJ Study as Originally Reported

The original study involved classes designated as fast, medium, and slow in reading at each grade level from first through sixth in a single elementary school, "Oak" School in South San Francisco. During May 1964, while Ss were in Grades K through 5, the "Harvard Test of Inflected Acquisition" was administered as part of a "Harvard-NSF Validity Study." As described to teachers, the new instrument purported to identify "bloomers" who would probably experience an unusual forward spurt in academic and intellectual performance during the following year. Actually, the measure was Flanagan's Tests of General Ability (TOGA), chosen as a nonlanguage group intelligence test providing verbal and reasoning subscores as well as a total IQ. TOGA was judged appropriate for the study because it would probably be unfamiliar to the teachers and because it offered three forms, for Grades K-2, 2-4, and 4-6, all of similar style and content. As school began in Fall 1964, a randomly chosen 20% of the Ss were designated as "spurters." Each of the 18 teachers received a list of from one to nine names, identifying those spurters who would be in his class. TOGA was then readministered in January 1965, May 1965, and May 1966.

RJ chose to obtain simple gain scores from the pretest (May 1964) to the "basic" posttest, a third testing in May 1965, and to make their primary comparisons with these. The main statistical computations were analyses of variance. Factors used in the analyses were treatment group (experimental vs. control), grade (first through sixth), ability track (fast, medium, slow), sex, and minority group status (Mexican vs. non-Mexican). An analysis of variance of the full 2x6x3x2x2 classification was neither planned nor possible since the experimental group contained only 20% of the children, only 17% of the total were Mexican, and the experiment was not designed to ensure equal representation by sex and ability track. Thus, with only 382 children actually included in the experiment, many of the 144 cells of the complete cross-classification table were empty (see our Table 2 for classroom by treatment group cell sizes). RJ calculated several two- and three-way analyses of variance using the unweighted means approximation to deal with problems of unequal cell frequencies.

The main results for Total IQ gain from pretest to basic posttest are presented in Chapter 7 of the RJ book. The main table of data is their Table 7-1, reproduced below, which shows mean gain in Total IQ for each grade and treatment group. "Expectancy advantage" was defined as mean gain for the experimental group minus mean gain for the corresponding control group (also called "excess of gain" by the experimental group). An excerpt from RJ's discussion follows:

> The bottom row of Table 7-1 gives the over-all results for Oak School. In the year of the experiment, the undesignated control-group children gained over eight IQ points while the experimental-group children, the special children, gained over twelve. The difference in gains could be ascribed to chance about 2 in 100 times (F = 6.35).
>
> The rest of Table 7-1 and Figure 7-1 show the gains by children of the two groups separately for each grade. We find increasing expectancy advantage as we go from the sixth to the first grade: the correlation between grade level and magnitude of expectancy advantage (r = -.86) was significant at the .03 level. (p. 74)†

The report continues with similar tables giving results for separate Reasoning and Verbal IQ scores and showing gain or "expectancy advantage" for breakdowns by sex and ability track. Brief profiles of a "magic dozen" of the experimental group children are also included, detailing their pre- and posttest IQ scores, along with anecdotal descriptions of each child. The overall results are interpreted as showing "... that teachers' favorable expectations can be responsible for gains in their pupil's IQs and, for the lower grades, that these gains can be quite dramatic" (p. 98).†

Also provided were supplemental analyses of data from the second and fourth TOGA administrations as well as graded achievement in various school subjects, teacher ratings of classroom behavior, and a substudy of general achievement test scores. Charts such as those reproduced in Figure 1 are given to illustrate "the process of blooming."

Table 1

Mean Gain in Total IQ After One Year by Experimental and Control-Group Children in Each of Six Grades (Reprinted from RJ, their table 7-1, p. 75)[†]

Grade	Control		Experimental		Expectancy Advantage	
	N	Gain	N	Gain	IQ Points	One-tail p < .05*
1	48	+12.0	7	+27.4	+15.4	.002
2	47	+ 7.0	12	+16.5	+ 9.5	.02
3	40	+ 5.0	14	+ 5.0	- 0.0	
4	49	+ 2.2	12	+ 5.6	+ 3.4	
5	26	+17.5(-)	9	+17.4(+)	- 0.0	
6	45	+10.7	11	+10.0	- 0.7	
Total	255	+8.42	65	+12.22	+ 3.80	.02

*Mean square within treatments within classrooms = 164.24

They show excess of IQ gain by experimental group over control group across testing occasions for various breakdowns of the school population.

The book concludes with a discussion of selected methodological criticisms of the study and more general methodological aspects of Hawthorne and expectancy studies, including design suggestions. It also offers speculation on possible processes of intentional and unintentional influence between teachers and students, and closes as follows:

> There are no experiments to show that a change in pupils' skin color will lead to improved intellectual performance. There is, however, the experiment described in this book to show that change in teacher expectation can lead to improved intellectual performance.

> Nothing was done directly for the disadvantaged child at Oak School. There was no crash program to improve his reading ability, no special lesson plan, no extra time for tutoring, no trips to museums or art galleries. There was only the belief that the children bore watching, that they had intellectual competencies that would in due course be revealed. What was done in our program of educational change was done directly for the teacher, only indirectly for her pupils. Perhaps, then, it is the teacher to whom we should direct more of our research attention. If we could learn how she is able to effect dramatic improvement in her pupils' competence without formal changes in her teaching methods, then we could teach other teachers to do the same. If further research shows that it is possible to select teachers whose untrained interactional style does for most of her pupils what our teachers did for the special children, it may be possible to combine sophisticated teacher selection and placement with teacher training to optimize the learning of all pupils.
>
> As teacher-training institutions begin to teach the possibility that teachers' expectations of their pupils' performance may serve as self-fulfilling prophecies, there may be a new expectancy created. The new expectancy may be that children can learn more than had been believed possible, an expectation held by many educational theorists, though for quite different reasons (for example, Bruner, 1960). The new expectancy, at the very least, will make it more difficult when they encounter the educationally disadvantaged for teachers to think, "Well, after all, what can you expect?" The man on the street may be permitted his opinions and prophecies of the unkempt children loitering in a dreary schoolyard. The teacher in the schoolroom may need to learn that those same prophecies within her may be fulfilled; she is no casual passerby. Perhaps Pygmalion in the classroom is more her role. (p. 182)†

Preview of Chapters 2-6

At this point, we give the reader a preview of the contents of the rest of the report. We have arranged our comments in five major sections: review of the RJ report, discussions of design and sampling problems, measurement problems, analysis problems and reanalysis results, summary and conclusions.

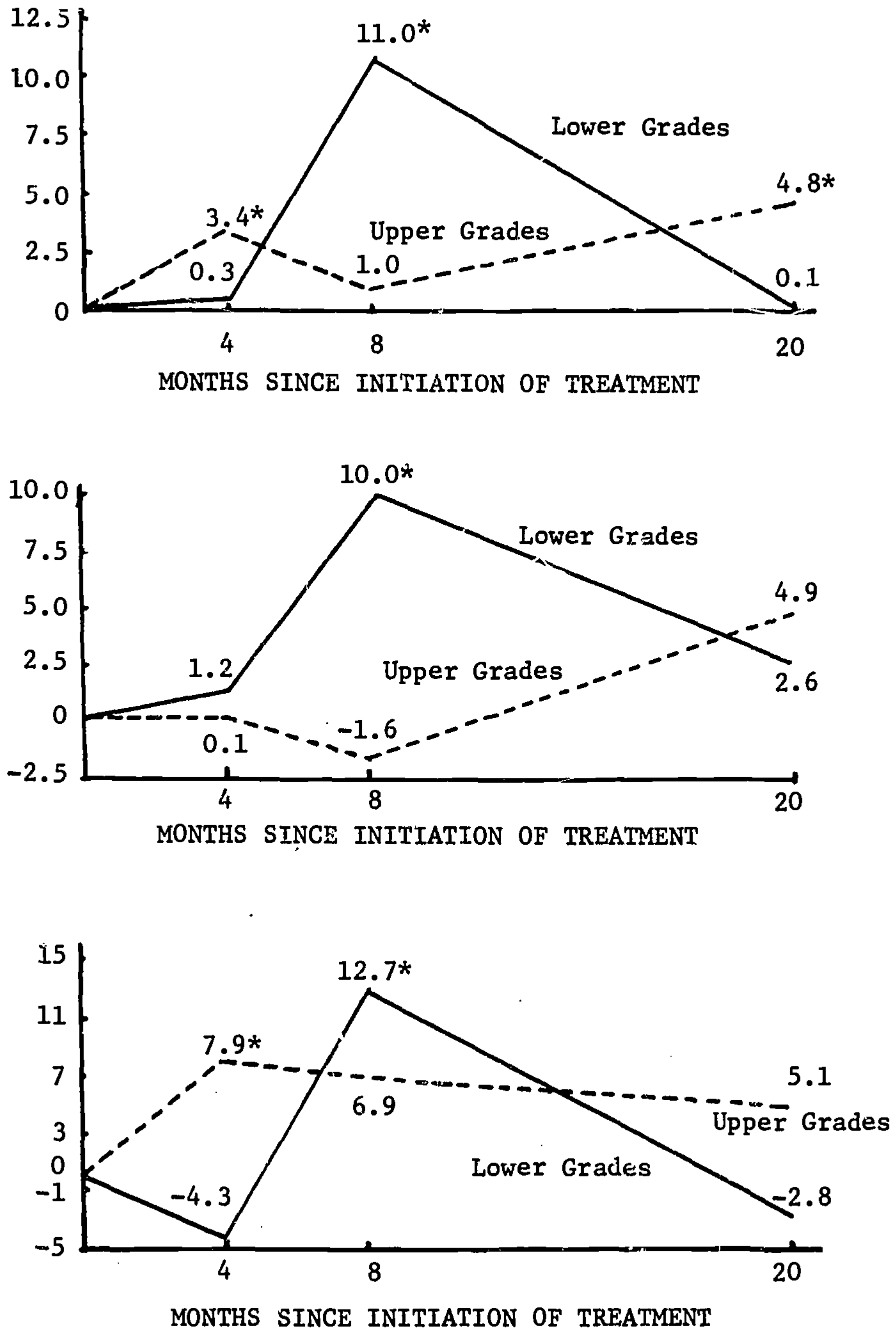


Figure 1: Expectancy Advantage After Four, Eight and Twenty Months Among Upper and Lower (Two) Grades (asterisk indicates $p < .10$ two-tail).
(Reprinted from RJ, their figure 9-5, p. 141.)[†]

The research report is a crucial part of the research process. Chapter II contains a critical review of Pygmalion as a research report. It is suggested that the report as a whole is inadequate. Descriptions of design, basic data, and analysis are incomplete. Inconsistencies between text and tables, overly dramatic conclusions, oversimplified, inaccurate or incorrect statistical discussions and analyses all contribute to a generally misleading impression of the study's results.

Chapter III examines RJ's experimental design and sampling procedures. The major difficulties discussed are the lack of clarity about the details of assignment to treatment groups, subject losses during the experiment, and the lack of balance in the design. These difficulties are especially important in the RJ study since the experimental group showed higher pretest scores on the average.

In Chapter IV, we examine the IQ scores actually obtained by children in Oak school, and questions of norming, reliability, and validity for these measurements. Histograms of the score distributions in each grade are shown. The number of IQ scores below 60 and above 160 especially for Verbal and Reasoning subscores raise doubts about the validity of the experiment as a whole and the results of certain statistical techniques in particular.

Chapter V contains a discussion of the methodological problems involved in the analysis of a complex study, comments on RJ's choice of analysis, and the results of our reanalyses. We demonstrate the wide variation in apparent results obtained from slightly different statistical approaches when serious imbalance in the design and major measurement problems exist.

Our overall conclusions about the results of the RJ study and some general methodological recommendations comprise Chapter VI.

Appendix A contains a glossary of terms and procedures referred to in the text. The raw data of the study are presented in Appendix B.

CHAPTER II: PYGMALION IN THE CLASSROOM AS A REPORT OF ORIGINAL RESEARCH

Before discussing methodological aspects of the RJ study, we consider it appropriate to examine the RJ book as a report of original research. A researcher's responsibility does not end when the experiment has been conducted and analyses concluded; he must report to the public his methods and findings. This is not a trivial final step but a crucial part of the research process. The benefits gained through careful experimentation may be lost if the final report is misleading. A careful reading of the report should provide the reader with sufficient information to allow replication of the study, to allow replication of the data analyses if provided with the data, and to allow him to draw his own conclusions about the results. Stated conclusions, tables, and charts should be carefully presented so that the uninformed reader will not be misled. All studies have weaknesses in design, execution, measurement, or analysis. These should be carefully discussed in the report because they affect the interpretation of results.

Careful reporting is especially important when the report receives considerable attention from methodologically unsophisticated readers, as in the case of Pygmalion. The phenomenon of teacher expectancy may be of central importance in the improvement of education, particularly if the scholastic development of disadvantaged children is strongly dependent on such effects. The problem then is of considerable social moment and the results of the RJ work have been widely distributed with noticeable impact in the news media. The following represents a sample of popular reaction:

> Can the child's performance in school be considered the result as much of what his teachers' attitudes are toward him as of his native intelligence or his attitude as a pupil? ... Pygmalion in the Classroom is full of charts and graphs and statistics and percentages and carefully weighed statements, but there are conclusions that have great significance for this nation.... Among the children of the first and second grades, those tagged "bloomers" made astonishing gains.... TOGA's putative prophecy was fulfilled so conclusively that even hard-line social scientists were startled. (Robert Coles, What Can You Expect?, The New Yorker, April 9, 1969, p. 172, 174);
>
> Here may lie the explanation of the effects of socio-economic status on schooling. Teachers of a higher socio-economic status expect pupils of a lower socio-economic status to fail (Robert Hutchins, Success in Schools, San Francisco Chronicle, August 11, 1968, p. 2);
>
> Jose, a Mexican American boy ... moved in a year from being classed as mentally retarded to above average. Another Mexican American child, Maria, moved ... from "slow learner" to "gifted child," .... The implications of these results will upset many school people, yet these are hard facts (Herbert Kohl, Review of Pygmalion in the Classroom, The New York Review of Books, September 12, 1968, p. 31);
>
> The findings raise some fundamental questions about teacher training. They also cast doubt on the wisdom of assigning children to classes according to presumed ability, which may only mire the lowest groups into self-confining ruts (Time, September 20, 1968, p. 62).

Other comments appeared in the Saturday Review (October 19, 1968) and a special issue of The Urban Review (September, 1968) was devoted solely to the topic of expectancy and contained a selection from Pygmalion. Rosenthal was even invited to discuss the results on NBC's "Today" show, thus reaching millions of viewers with the idea. The study was also cited in at least one city's decision to ban the use of IQ tests in primary grades:

> The Board of Education's unanimous action was founded largely on recent findings which show that in many cases the classroom performance of children is based on the expectations of teachers.
>
> In one study conducted by Robert Rosenthal of Harvard University, the test results given to teachers were rigged, but the children performed just as teachers had been led to expect based on the IQ scores. (Jack McCurdy, Los Angeles Times, January 31, 1969)

Because the book received wide attention and will likely stimulate more public discussion and policy decisions as well as much further research, it is imperative that its results be thoroughly evaluated and understood. Unfortunately, a complete understanding of the data and results are not obtainable from the published accounts alone.

Pygmalion in the Classroom can be severely criticized as a research report. We summarize our criticisms briefly here and then return to each in more detail. The RJ report is misleading. The text and tables are inconsistent, conclusions are overdramatized, and variables are given prejudicial labels. The three concluding chapters represent only superficial, and frequently inaccurate, attempts to deal with the study's flaws. Descriptions of design, basic data, and analysis are incomplete. The sampling plan is not spelled out in detail. Frequency distributions are lacking for either raw or IQ scores. Comparisons between text and appendix tables are hampered by the use of different subgroupings of the data and the absence of intermediate analysis-of-variance tables. Many tables and graphs show only differences between difference scores, i.e., gain for the experimental group minus gain for the control group. There are technical inaccuracies: charts and graphs are frequently drawn in a misleading way and the p-value or significance level is incorrectly defined and used. Statistical discussions are frequently oversimplified or completely incorrect (some of the statistical questions are considered in later sections).

In short, our criticisms can be stated in the more general words of Huff (1954):

> The fault is in the filtering-down process from the researcher through the sensational or ill-informed writer to the reader who fails to miss the figures that have disappeared in the process.

Interpretations and Conclusions

Conclusions are frequently overstated and do not always agree from place to place in the book. Text and tables are not always in agreement. Again, our concern is well stated by Huff (1954, p. 131):

> When assaying a statistic, watch out for a switch somewhere between the raw figure and the conclusion. One thing is all too often reported as another.

RJ use labels for their dependent variables that presume interpretations before effects are found, a practice especially to be condemned in publications aimed at the general public. "Intellectual growth" is used in referring to the simple difference between a child's pretest IQ score and his IQ score on a posttest. It is questionable whether simple gain from first to a later testing (with some adjustments for age) using the same test represents anything so global as intellectual growth.

The difference in gains shown by the experimental group over the control group is described as an "expectancy advantage." This term presupposes that the difference is always positive. In fact it is not. What particular "advantage" or "benefit" accrues to the child showing a large gain score is not made clear. Words like "special" and "magic" are also frequently used to refer to experimental children, when less provocative terms would serve as well.

Looking at RJ's main results for Total IQ, as reported in their table 7-1 (see our Table 1), the 1st and 2nd grade experimental groups show a large significant expectancy advantage, the 4th graders show a small nonsignificant advantage, the 3rd and 5th graders show no difference and the 6th graders show a small nonsignificant disadvantage. So RJ's table reports an "expectancy advantage" for the first and second graders (and possibly the 4th graders) and reports no "expectancy advantage" for the other grades. The significant "expectancy advantage" reported by RJ is thus based only on the 19 first and second graders in the experimental group. But RJ conclude:

> We find increasing expectancy advantage as we go from the sixth to the first grade.... (p. 74) †

Here is how RJ describe the results elsewhere in the text:

> When the entire school benefitted as in Total IQ and Reasoning IQ, all three tracks benefitted. (p. 78)
>
> When teachers expected that certain children would show greater intellectual development, these children did show greater intellectual development. (p. 82)
>
> The evidence presented in the last two chapters suggests rather strongly that children who are expected by their teachers to gain intellectually in fact do show greater intellectual gains after one year than do children of whom such gains are not expected. (p. 121) †

> After the first year of the experiment a significant expectancy advantage was found, and it was especially great among children of the first and second grades. (p. 176)†

There is thus a clear tendency to overgeneralize the findings. When the authors are explaining away the results of contradictory experiments, however, the conclusions sound quite different:

> The finding that only the younger children profitted after one year from their teachers' favorable expectations helps us to understand better the [negative] results of two other experimenters.... (p. 84)†
>
> The results of our own study suggest that after one year, fifth graders may not show the effects of teacher expectations though first and second graders do. (p. 84)†

Another important inconsistency is between the form of analysis and the stated conclusions. All analyses were done in terms of means, yet conclusions are stated in terms of individuals; for example "... when the entire school benefitted...." or "...these children did show greater intellectual development." That is, the analyses performed by RJ could only show that average gains by experimental children were larger than average gains by control children, but RJ's statements imply that each individual experimental child gained and that these gains were all larger than those shown by any control group child.

There is a strong presumption throughout the book that teacher expectations have an effect. Contrary evidence is explained away. RJ cite other studies which in general did not support the conclusions drawn in this book. The discussion of these adverse findings de-emphasizes the possibility that teacher expectations have little effect on IQ scores and becomes almost absurd with references to all possible alternative hypotheses--"there is such an effect, but..." (RJ, p. 57).†

One of RJ's closing chapters takes steps toward answering specific methodological criticisms. Unfortunately, much of this discussion is superficial and some is incorrect. (See later chapters on technical inaccuracies, design and sampling, and reliability.) RJ's chapter also offers speculation on possible processes of intentional and unintentional influence between the teachers and students, but fails to face the full implications of the fact that after the study the teachers could not remember the names on the original lists of "bloomers" and reported having scarcely glanced at the list.

RJ's last chapter provides a capsule summary and some general implications. It is here that the inadequacy of statistical summaries of these data should be clearly specified. But it is not. The reader expecting careful conclusions is given overdramatized generalities instead.

Tables, Figures and Charts

Even with a faulty text, a reader should be able to examine the basic figures, tables, and analyses and draw his own conclusions. Clearly in a massive study, we cannot demand that an author include all the data, or a complete set of analysis-of-variance tables, etc. RJ indeed included many appendix tables of summary data. What then is wrong?

Nowhere can the reader see the distributions of pretest or posttest scores, the relationship between pretest and posttest scores, or the detailed results of any of the analyses. The tables in the body of the text show mean gain or "excess of gain" from pretest to posttest for treatment groups in breakdowns by grade, sex, track, or some combination of factors. Excess of gain is mean gain by the experimental group minus mean gain by the corresponding control group. This obscures the fact that some of the startling gains are made by children whose pretest IQs were far below reasonable levels for normal school children. Examination of alternative hypotheses, such as "that children higher (or lower) to begin with gain more," or "that unreliability may have contributed to spurious results," are hampered. Means and standard deviations for pretest, posttest, and gain are shown in the appendix but not for the same breakdowns as shown in the text. Selected means or standard deviations to compare with text tables, such as Table 7-1 which shows a breakdown by grade, can be obtained with some computation. But for RJ tables such as Table 7-5 showing breakdown by sex, it is impossible to obtain mean pretest or posttest scores from data supplied in the book. Since no analysis-of-variance tables are shown, the reader must rely on statements like "The interaction term was not very significant (p < .15)...." (RJ, p. 77).[†] However, there were several analyses of variance, with different combinations of factors yielding different results, so p values quoted in the text were all obtained from different analysis of variance calculations. The reader is left uncertain as to which results were obtained in what analysis and cannot reconstruct tables of means to interpret each effect for himself.

Since final interpretations of the results and the validity of many of the statistical procedures RJ employed rests on the score distributions and the relationships of pre to post scores, the reader would hope to find tables, histograms, and scatterplots to enable him to examine the data more closely, at least for the main subsets of data. At the very least, the authors should be able to assure the reader that they have examined the data in this light and are satisfied. But no histograms or frequency distributions of individual scores are provided or mentioned. If these were displayed, the reader would notice that Total IQ scores range from 39 to 202, Reasoning IQ scores range from 0 to 262, and Verbal IQ scores range from 46 to 300. (See Chapter IV for a discussion of the meaning of extreme scores like these.) There are also no scatterplots showing relationships between pretest and posttest scores.

Of the nine figures in RJ Chapters 7-9, eight are drawn in a misleading way; Huff calls graphs like these "gee-whiz" graphs. RJ Figure 7-2, which also appeared in Scientific American (RJ, 1968a), is mislabelled, does not state that its impressive percentages are based on a total of only 19 children in the experimental group, with the 4 children gaining 30 or more points included with those gaining 20 or more points who in turn have been included with the children gaining 10 or more points. Our Figure 2b shows the information in RJ's Figure 7-2 redrawn to eliminate overlapping or repetition of information and inaccurate labelling.

RJ Figures 8-1, 8-2, 9-1, and 9-2 all are drawn with false zero lines, over-emphasizing apparent gains and differences in gain. For example, in RJ Figure 8-1 the line of zero gain is in the middle of the chart and the entire scale displayed on the graph runs from -0.5 to +0.8 grade points based on a scale from 0 for "F" to 4 for "A". The choice of scale makes the gains and differences in gains look large when, in fact, most are considerably less than one gradepoint. Our Figures 3a and 3b show RJ's Figure 8-1 and the same figure redrawn appropriately.

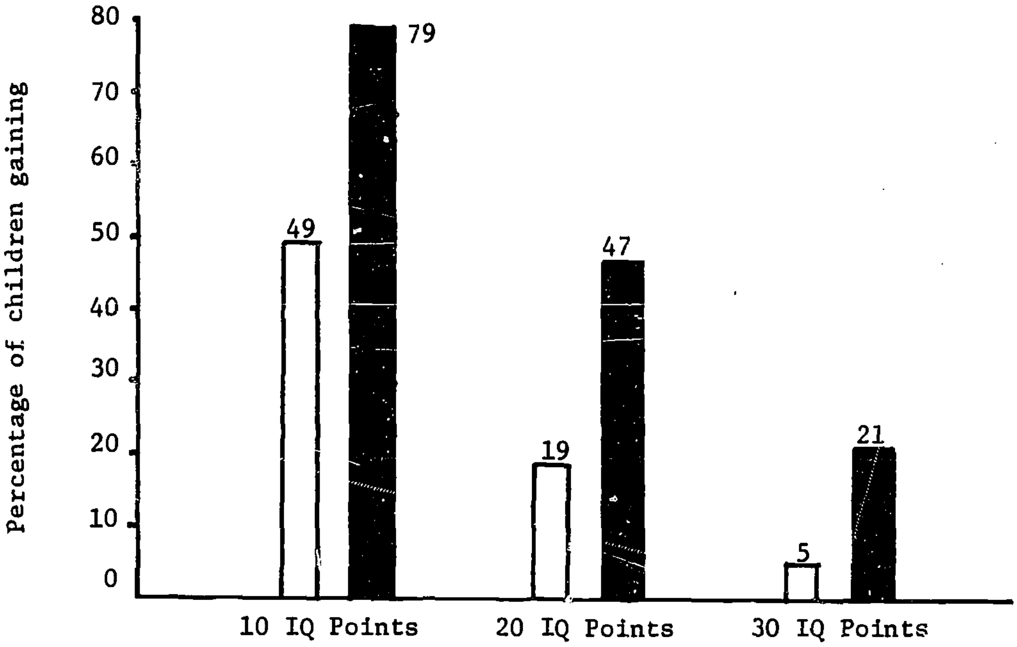
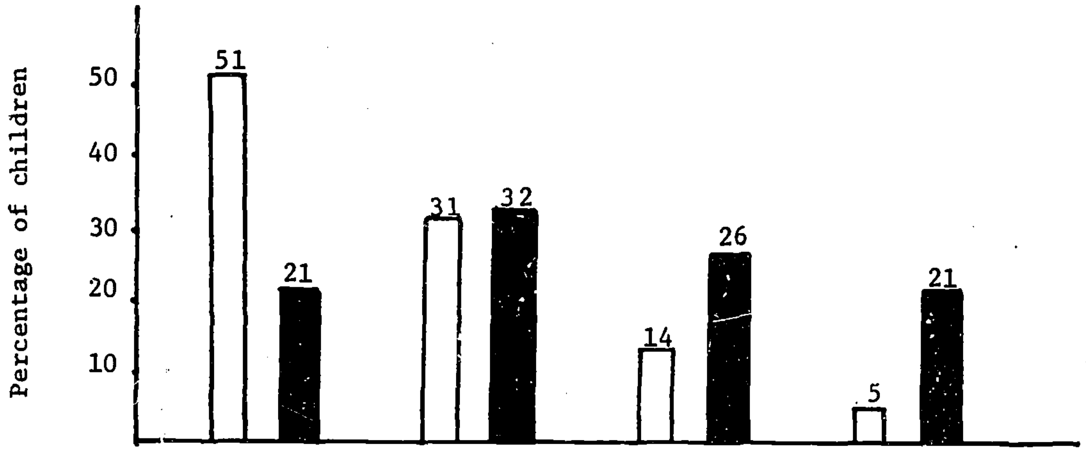


Figure 2a: Percentages of First and Second Graders Gaining Ten, Twenty, or Thirty Total IQ Points
(Reprinted from RJ, their figure 7-2, p. 76)†

□ Control Group

■ Experimental Group



Gain in IQ pts.	G < 10		10 < G < 20		20 < G < 30		G > 30		Total	
No. of children	48	4	29	6	13	5	5	4	95	19
% of children	50.5	21	30.5	32	14	26	5	21		

Figure 2b: RJ Figure 7-2 Redrawn to Eliminate Repetition (Note that "gains" actually varied from -17 to +65)

□ Control Group

■ Experimental Group

Figures such as 8-1, 8-2, 9-1 and 9-2 should be drawn with the zero line strongly indicated and all gains originating from it.

The four "process of blooming" charts (RJ Figures 9-3 through 9-6) not only display floating zero lines and elastic scales from one IQ measure to another, but particular measures are drawn on different scales in each chart so that comparisons between charts are not possible. (Scales for the IQ differences are 0 to 5, -3 to 12, 0 to 12.5, and 0 to 6 respectively.) More important, the "expectancy advantage" computed at each time point is based on a different set of children, since there are missing data and subject losses along the way. Finally, all the charts indicate no "expectancy advantage" at Time 1 (the pretest). Since the experiment had not begun there are no gains to compare, but in fact the two groups did not have the same average pretest scores. For example, for the Total IQ chart in Figure 1 the experimental group had average pretest scores 4.9 IQ points higher than the control group in the lower grades and 2.4 IQ points higher in the upper grades (these numbers obtained from our Table 20 in the re-analysis chapter).

Technical Inaccuracies

Books intended for use by students should be free from technical inaccuracies. One striking deficiency here is RJ's misuse of p-value. The concept of p-value or significance level is incorrectly defined and interpreted throughout the book. In the preface, p-value is defined incorrectly:

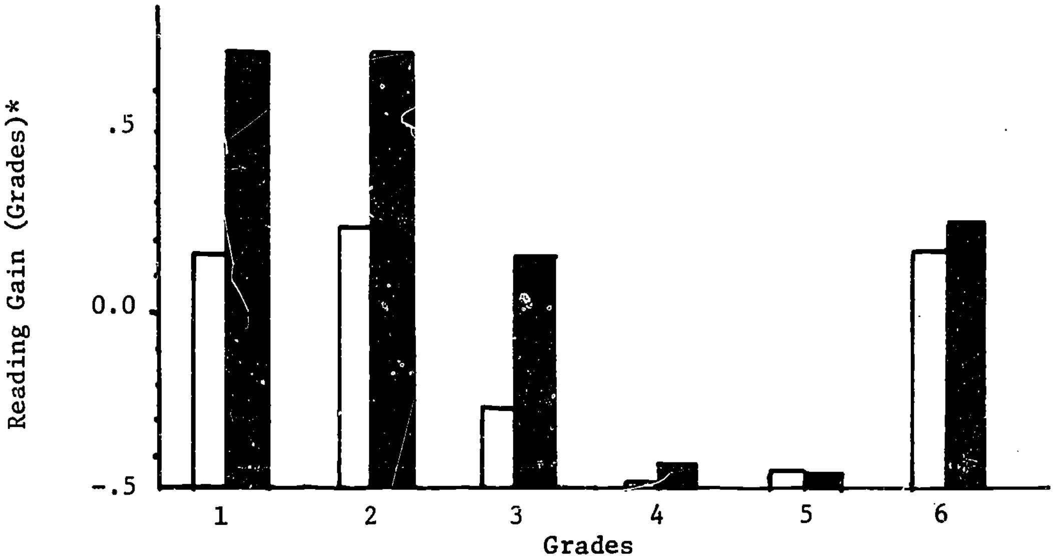


Figure 3a: Gains in Reading Grades in Six Grades (Reprinted from RJ, their figure 8-1, p. 100)†

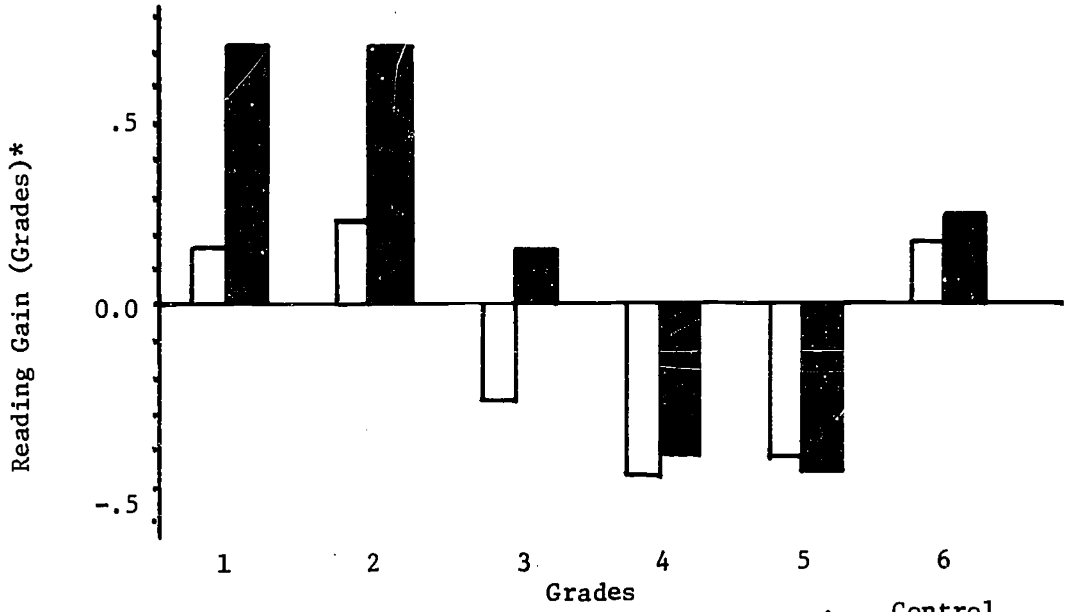


Figure 3b: RJ Figure 8-1 Redrawn with Gains Beginning at Zero

*These are coded teacher's marks (A = 4, B = 3, C = 2, D = 1, F or U = 0), not grade equivalent scores

> "... there often will be a letter p with some decimal value, usually .05 or .01 or .001. These decimals give the probability that the finding reported could have occurred by chance. For example, in comparing two groups the statistical significance of the difference in scores may be reported as t = 2.50, p < .01, one-tailed. This means that the likelihood was less than 1 in 100 that the difference found could have occurred by chance." (p. ix)†

This definition should read: this means that the likelihood was less than 1 in 100 that the difference found *or one larger* could have occurred by chance *under the null hypothesis that the true difference was zero.* The trouble with RJ's definition is its implication that the observed difference *is* the true difference, that because this particular difference is unlikely to have occurred by chance it must be real. The definition also ignores the fact that this p-value can only be determined *if* certain assumptions are true. The p-value does not tell us how close an observed difference is likely to be to the true difference. It simply identifies the likelihood of a more extreme result than the one observed *given that the null hypothesis is true*. For example, if a *t*-test based on a difference in sample means of, say, 10.2 yields p < .01, one-tail, this means that the probability of observing a difference in sample means as large or larger than 10.2 is less than .01 *if in fact there is no real difference in population means and all the assumptions necessary for the test to be valid are satisfied*. The "true difference" need not be anywhere near 10.2. For example, the probability of observing a difference in sample means by chance more extreme than 10.2 if the "true difference" were 6.8 is about .22.

RJ seem also to use p-value as a measure of strength of effect, an indication of the size and practical importance of mean differences.

They do not use a standard p-value such as .05, preferring to quote values ranging from .25 to .00002 thus encouraging the reader to conclude that p-values of .001 indicate truer, larger, more important effects than p-values of .01. The p-value is not a useful measure of the size or importance of an observed treatment effect for individuals because it depends on the sample sizes involved as well as the actual size of the difference. Small differences of no practical importance can be shown statistically significant at a small p-value if the sample size is large enough. Conversely, large differences may not be statistically significant if the sample size is small. Procedures which can be used to assess the size of treatment effects include: confidence interval for the differences in means, histograms showing the relative positions of control group scores and experimental group scores, percent of individuals misclassified, measures of statistical association such as ω^2 (Hays, 1963), and linear regression analysis showing the percent of variance accounted for by treatment relative to other factors.

Most importantly, however, it is usually meaningless to quote particular p-values less than .01 since the actual distribution of a statistic such as _t_ in a real problem will seldom be well approximated by the tabled distribution far enough into the tails (see our later section on reliability) for small p-values to be meaningful.

RJ devote nine pages to a discussion of the higher gains in reading grades shown by the experimental or "special" children. Yet they state:

> When the entire school was considered, there was only one of the eleven school subjects in which there was a significant difference between the grade-point gains shown by the special children and the control-group children. (p. 99)†

Why is so much emphasis placed on results for one out of eleven school subjects? A series of eleven independent _t_-tests at the 10% level referred to by RJ can be expected to produce at least one significant difference by chance even though there is no true difference in any of the eleven. In fact, the probability of obtaining at least one significant difference by chance under these circumstances is .6862*. Of course, these sets of grades are not independent and the probability of obtaining at least one significant result by chance will be smaller than .6862 but will undoubtedly be considerably larger than .10.

In a footnote, RJ argue that:

> Even allowing for the fact that reading was the only school subject to reach a $p < .10$ of a total of eleven school subjects, these obtained p's for reading seem too low to justify our ascribing them to chance. If the eleven subjects were independent, which they were not ... we might expect on the average to find by chance one $p < .09$, and that expected p is about ten times larger than those obtained when classrooms served as sampling units. (p. 118-119)†

The problem of "expected p-values" needs further examination. First, no matter how small the p-value is, the difference may not be real; there is always the chance that a rare event has occurred. Second, what is the probability of a very small p-value given that the p-value is less than .10? It is easiest to examine this question for the sign test on seventeen classes, for which the obtained p-value for reading scores was .0062. Given that $p < .10$, and that the probability of $E > C$ is one

*$P(\underline{t} \text{ significant } |H_0) = .10$, $P(\text{no } \underline{t} \text{ significant } |H_0, 11 \text{ independent } \underline{t}\text{'s}) = (.90)^{11}$, $P(\text{one or more } \underline{t} \text{ significant } |H_0) = 1-(.90)^{11} = .6862$.

half, the probability that the p-value is less than or equal to .0062 is .0879. In other words, there is about a 9% chance of a p-value as small or smaller than .0062 given that $p < .10$. In such circumstances, a confidence interval for the difference in reading scores would provide more information about the practical importance of obtained results than any discussion of p-value.

CHAPTER III: DESIGN AND SAMPLING PROBLEMS

There are several problems inherent in the design of the RJ study and the sample finally obtained. We list them briefly and then discuss each in turn. The sampling plan, the procedure for assignment of children to treatment groups, is ill-defined. Little balance was designed into the study. A 20% subject loss from pretest to posttest reduces the generalizability of the study and raises the possibility of differential subject loss in experimental and control groups. Because of the uncertain sampling plan, the lack of balance and the possibility of non-random subject loss during the experiment, the fact that the experimental group showed higher pretest scores on the average, especially in the lower grades, suggests serious difficulties that attempts at statistical correction may not erase.

The details of a sampling plan provide the basis for subsequent statistical inference as well as for planning replications of a study. In addition, the sampling plan determines the population to which the results can be generalized, the unit of observation (individual or classroom), the comparability of experimental and control groups, and the factors which may be used in an analysis of variance. It is not clear from the RJ book just what the procedure for assignment to treatment groups was. According to the authors, a 20% random sample of the school's children were listed as "bloomers" to form the experimental group. However, "... it was felt to be more plausible if each teacher did not have exactly the same number or percentage of her class listed" (p. 70).† Thus, the number of experimental children in a classroom varied from one to nine. "For the same reason the proportion of either boys or girls on each teacher's list was allowed to vary from a minimum of 40 percent of the designated children to a maximum of 60 percent of the designated children" (p. 71).† Was this plan simple random sampling, or random sampling stratified by sex and classroom, or some compromise solution? It makes a difference in our choice of analysis. Perhaps simple randomization was followed by a nonrandom reassignment procedure to fit specifications; the authors do not say. In the final analysis do we actually have random assignment to treatments?

The major difficulty with the RJ design is the imbalance deliberately created to make the experimental condition plausible for the teachers. With highly variable human subjects and a small experimental group, it is especially important that the experimental and control groups be comparable on as many factors as possible. Statistical inference at the end of the experiment will rest on the finding that the experimental and control groups differ by more than could be expected on the basis of inherent variability. If groups differ for reasons other than the experimental treatment variable, results may be confounded and interpretation rendered impossible. A main objective of experimental design is to control sources of variability so that no confounding impedes interpretation.

As a result of subject loss during the experiment as well as original inequalities, the number of children in each classroom and treatment group available for the basic posttest varies as shown in Table 2. The percent of children in the experimental group from each classroom is also shown. The lack of equality in the number of experimental children per classroom means that some classes have too few experimental children

TABLE 2

Number of Children Taking the Basic Posttest

by Classroom and Treatment Group

		Track		
		Fast	Medium	Slow
Grade	Group			
1	C	17	15	16
	E	1 (6%)	4 (21%)	2 (11%)
2	C	19	14	14
	E	6 (24%)	3 (18%)	3 (18%)
3	C	12	15	13
	E	8 (40%)	1 (6%)	5 (28%)
4	C	18	16	15
	E	5 (22%)	3 (16%)	4 (21%)
5	C	16	-	10
	E	5 (24%)	-	4 (29%)
6	C	20	13	12
	E	4 (17%)	4 (24%)	3 (20%)
All Grades	C	102	73	80
	E	29 (22%)	15 (17%)	21 (21%)

to make analysis within classrooms feasible. The inclusion of sex as a factor in the analysis immediately creates empty cells. To counteract this, RJ combined other factors to do ANOVAS on treatment by sex, and treatment by sex by grade, for example, which necessitates combining over tracks and introduces confounding. Thus in the first grade, the experimental group comes mainly from the middle track while in the third grade the middle track is hardly represented at all; tracks are much more evenly represented in the control group.

In designing experiments like the one under discussion here, an appropriate procedure is first to match or block subjects on potentially important variables, like grade, sex, and classroom, and then to rely on random assignment of subjects to treatments within blocks to provide balance for other variables. This procedure insures that the groups are comparable on the blocked variables and thus equally representative of the population of interest. It is also advisable to check the adequacy of obtained balance in the subjects remaining in the experiment at the end; different experimental treatments can create differential dropout or loss rates among subjects, and this effect may dictate changes in the statistical analysis, as well as being of interest in its own right. Variables which have not been used in blocking may be included as factors in an analysis of variance only with considerable caution (see section on analysis of variance in unbalanced designs).

The plausibility of the lists of children expected to "bloom" is a crucial issue in an experiment of this type, but randomization and balance are also important. RJ could have taken some steps to achieve balance without giving every teacher a list including exactly the same number of names. The most important factor for balancing is perhaps ability track. Track assignments were made on the basis of reading ability by the previous year's teacher, after the administration of the TOGA pretest but without knowledge of these pretest IQs. There were three classes, representing the three tracks, at each grade level. Since classes apparently differed in size, assigning exactly the same proportion of children in each class would not have resulted in the same number of children on each list. If class size represented on the pretest is indicative of the whole experiment, total class size varied from 16 to 27; 20% of these classes would vary from three to five or six. It is questionable whether a teacher would notice that three in a class of 17 represents the same proportion as six in a class of 28. However, another possibility would have been to take a lower percentage of children from the fast track and a higher percentage of children from the slow track, since fast track children might be said to have already "bloomed." If all classes were of size 20, we might choose 15%, 20%, 25%, or three, four, and five experimental children in the fast, medium, and slow tracks, respectively. With such a small experimental group it is difficult to achieve balance on sex also, but perhaps teachers could be told that the prediction is done separately for the two sexes so the lists contain equal numbers of boys and girls. There seems little reason for allowing the number of experimental children in a class to vary haphazardly from one to nine. When many children are lost to the experiment through attrition, the original balance may be partially lost, but this is no reason to ignore the question of balance at the beginning.

There is the possibility of a selection bias of unknown proportions. Although 478 children were given the pretest, only 382 or 80% were present for at least one posttest and were thus "included in the experiment" (see Table 3). RJ remark that "The ins and outs seldom belong to the high or top-achieving third of the school" (p. 63).† Thus the children remaining in the experiment cannot be considered a random sample of Oak School children and the results may not be representative of the reactions of the whole school population. In view of the high subject loss, it is doubtful that the experimental and control children can still be regarded as representing comparable groups. Although roughly the same proportion of experimental and control children were lost to the experiment, pretest scores on lost subjects were not available and it is impossible to tell whether both groups lost comparable children.

Given the uncertain sampling plan and large subject loss, it is disconcerting to note that, for those children remaining in the experiment, the pretest scores are consistently superior in the experimental group.

TABLE 3

Number of Children Taking Pretest and at Least One Posttest

	Pretest only	Pretest and at least one posttest	Total pretested
Control	79	305	384
Experimental	17	77	94
Total	96	382	478

> In spite of random allocation to the experimental condition, the children of the experimental group scored slightly higher in pretest IQ than did the children of the control group. This fact suggested the possibility that those children who were brighter to begin with might have shown the greater gains in intellectual performance. (p. 150)†

In Chapter 10, RJ explore this possibility using two different procedures: one involves correlations between pretest scores and gain scores; the second is based on post hoc matching of experimental and control children. They conclude:

> These analyses suggest that the over-all significant effects of teachers' favorable expectations cannot be attributed to differences between the experimental- and control-group children in pretest IQ. (p. 151)†

But neither RJ procedure provides an adequate investigation of the possibility that children higher to begin with gained more. The correlation analysis is, in fact, incorrect. RJ state:

> As one check on this hypothesis, the correlations were computed between children's initial pretest IQ scores and the magnitude of their gains in IQ after one year. If those who were brighter to begin with showed greater gains in IQ, the correlations would be positive. In general, the over-all correlations were negative; for total IQ $r = -.23$ ($p < .001$); for verbal IQ $r = -.04$ (not significant); and for reasoning IQ, $r = -.48$ ($p < .001$). (p. 150)†

Actually, the correlation between pretest scores and gain scores can generally be expected to be negative. If X_i represents the pretest scores, and Y_i the posttest scores; their variances are σ_X^2 and σ_Y^2, and their correlation is ρ. Then the correlation between gain scores, $Y_i - X_i$ and pretest scores X_i is

$$\rho_{Y-X,X} = \frac{\rho\sigma_Y - \sigma_X}{\sqrt{(\sigma_Y - \sigma_X)^2 + 2(1-\rho)\sigma_X\sigma_Y}} \; .$$

Thus, $\rho_{Y-X,X}$ can be positive only if $\rho > \sigma_X/\sigma_Y$. Since σ_X/σ_Y should seldom be much smaller than 1.0, we see that the correlation between gain scores and pretest scores will generally be negative. (If, for example $\sigma_X = \sigma_Y$ and $\rho = .68$ which is a situation representative of the RJ data, see Tables 4, 5, 6, then $\rho_{Y-X,X} = -.4$).

Clearly, correlations between pretest scores and gain scores are determined by the correlation between pretest scores and posttest scores and cannot be used to investigate whether those who were brighter to begin with gained more. If pretest and posttest scores have a linear relationship and those with higher pretest scores gain more, the slope (β) of the regression equation of posttest on pretest will be greater than unity. If those with higher pretest scores gained a great deal more, one might expect to find a nonlinear relationship between pre and posttest. Referring to our reanalysis section, note that the slope is generally less than unity although it is larger than unity for grades 5 and 6 Total and Verbal IQ and grades 3 and 4 Verbal IQ (Tables 9 and 10). Note however, that Figures 11 through 19 show nonlinear effects produced by a few children with high pretest scores and large gains.

RJ's second procedure was to match experimental and control group children within classrooms on pretest scores and to compute an "expectancy advantage" for each matched pair. Post hoc matching can be useful only when close objectively chosen matches are possible. Since the experimental group was only 1/4 the size of the control group, choosing a control child to match each experimental child must involve subjective decisions. Also, the fact that 13 of the 65 experimental children were left unmatched indicates a lack of comparability of the two groups. Our reanalysis section presents some further evidence on the difficulties involved in post-hoc matching.

CHAPTER IV: MEASUREMENT PROBLEMS

For the main purposes of their study, RJ chose TOGA, a group intelligence test which purportedly does not require reading ability. RJ obtained individual IQ scores for each testing and defined changes in these scores over time as "intellectual growth." TOGA forms K-2, 2-4, and 4-6 were used. On the pretest K-2 was administered to the kindergarten and first grade classes, form 2-4 was administered to the second and third grades, and form 4-6 was administered to the fourth and fifth grades. On the second and third tests during the following year all children were retested with the same test form (grade designation used by RJ was that at basic posttest). On the fourth test, two years after the pretest, those who had been in kindergarten, second grade, or fourth grade on the pretest were again tested with the same TOGA form while the other children were tested with the next-higher-level form. These IQ tests were multiple choice with 5 choices for each item, forms K-2 and 2-4 each had 63 items, 35 verbal and 28 reasoning, form 4-6 had 85 items. Thus for example, children in kindergarten on the pretest, first grade for second and third tests, and second grade for the fourth test received form K-2 all four times, while children in the first grade on the pretest, second grade for the basic posttest, and third grade for the last test received form K-2 the first three times and form 2-4 for the fourth time.

Among the most important questions to be asked, here as in any research project, are: What is being measured? How is it being measured? How accurately is it being measured? What scale of measurement is being used? In this section we examine the IQ scores actually obtained by children in Oak school, and questions of norming, reliability, and validity for these measurements.

Scores and Norms

Problems began with the decision to rely solely on TOGA. Examination of the manual suggests that the test has not been fully normed for the youngest children, especially for children from lower socio-economic backgrounds. In addition, it was administered to separate classes by the teachers themselves, a fact which raises doubts about standardization of procedure. A review of the test manual shows that for grades K-2 the procedure is regarded more as a class project than as a test. Although the teacher reads each item in the verbal subtest, in the reasoning portion children are left on their own with only minimal instruction or guidance from the teacher. There appears to have been no attempt to train the teachers in test administration, to check the adequacy of administration, or to determine whether the test and its instructions and procedure were understood by the subjects. With kindergarteners and first graders, in particular, it is doubtful that any closely timed group test can be regarded as an adequate measure of intellectual status.

All computations were based on IQ scores--a transformation of the raw scores based on norm groups and the age of the child. The total raw score distribution on form K-2 for example has a possible range of 0 to 63 points. Examining the conversion table, one notes that a difference in raw scores of one item on TOGA will result in an IQ difference (for children of the same age) of about 2 points near the center of the distribution, up to 8 points at the bottom of the scale, and 60 points at the top.

According to the manual, TOGA IQ scores were normed so that for school children the mean IQ should be 100 (although it might be lower for some socio-economic groups) and the standard deviation should be 16 or 17. Thus 95% of the children should be in the range 67 to 133. A detailed table of mental ages corresponding to each raw score from one to the maximum possible is provided in the manual. In a technical report accompanying TOGA, norms showing mental age extrapolated up to 26.6 and down to zero are provided. As Thorndike (1969) notes elsewhere, however, extrapolations outside the norm sample range are of questionable value. However, the tables showing IQ scores for each raw score and age are not extrapolated beyond IQs of 60 and 160. Thus although it is possible to obtain IQs of 0 to 200 or more using information provided in the manual, the manual implicitly discourages use of IQs lower than 60 or higher than 160, which should occur very rarely in any case.

One simple check on the adequacy of the IQ scores provided by TOGA would be a comparison of the score distribution obtained for the "Oak School" children with those of the norming groups. RJ provide no score distributions in either text or appendix, although examining RJ tables A-1, A-2, and A-3 in the appendix we find pretest Total IQ means within treatment group of 60.5, 76.9, 79.9 for some low track classrooms. The pretest mean for Reasoning IQ was 58.0 for the entire first grade; in the first grade control group, Reasoning pretest means were 30.8 and 47.2 for slow and medium track, respectively. It should be noted that, at one time, children with IQs below 70 were officially described as feeble-minded. Those below 40 were labeled "imbeciles." Today, a score of 75 or below usually identifies individuals for special EMR* classes. Since IQ scores

*Educable mentally retarded

as high as 60 could easily be obtained by "guessing" on form K-2 (see below) IQ scores as low as these must include random or systematically incorrect responses and unattempted items (an IQ of 63 for a 6 year old represents 12 correct out of 63 multiple choice items). Some IQ means seemed inconsistent with the tracking classification; for the third grade control group, fast, medium, and slow track pretest total IQ means were 98.4, 102.2, 100.3 respectively. Pretest means for different forms of TOGA also seemed inconsistent; first and second graders had a mean total IQ of 92.3, third and fourth graders of 104.3 and fifth and sixth graders of 99.2.

As a consequence, our first step was to examine the score distributions in detail. Histograms of Total IQ, Verbal IQ, and Reasoning IQ scores on pretest and basic posttest for each grade are shown in Figures 4-9. Means, standard deviations, and maximum and minimum scores are shown in Tables 4 and 5.

Notice the pretest Reasoning IQs of zero in the first grade (Figure 6), the posttest Total IQs of 202 in the second grade, the posttest Verbal IQs of 221, 249, 300, the posttest Reasoning IQs of 251, 262.

Since Total IQ scores on the pretest were so low for first and second graders, it is interesting to compare the obtained distribution with that to be expected if children merely "guessed." TOGA form K-2 is a multiple choice test with five choices for each of 63 items. If we define "guessing" to mean that a child selects at random one of the five choices and each choice is made with probability 1/5, then raw scores on

Table 4

Pretest Scores

All Pretested Children with at Least One Posttest

Grade	N	Mean	Standard Deviation	Minimum	Maximum
			Total IQ		
1	63	90.0	19.4	39	130
2	63	94.7	15.8	59	133
1 & 2	126	92.3	17.9	39	133
3 & 4	131	104.3	17.4	64	158
5 & 6	125	99.2	18.4	56	152
			Reasoning IQ		
1	63	58.0	36.8	0	111
2	63	89.1	21.6	39	133
1 & 2	126	73.5	34.1	0	133
3 & 4	131	99.5	19.5	56	167
5 & 6	125	96.6	20.3	52	158
			Verbal IQ		
1	63	105.7	21.2	54	183
2	63	99.4	16.1	50	133
1 & 2	126	102.6	19.2	50	183
3 & 4	131	109.7	22.2	68	171
5 & 6	125	102.6	24.4	46	165

Table 5

Basic Posttest Scores

Grade	N	Mean	Standard Deviation	Minimum	Maximum
			Total IQ		
1 & 2	114	103.4	18.4	67	202
3 & 4	115	107.7	20.1	57	165
5 & 6	91	112.3	22.8	63	171
			Reasoning IQ		
1 & 2	114	102.3	29.2	39	211
3 & 4	115	103.6	28.5	0	262
5 & 6	91	116.5	29.7	67	251
			Verbal IQ		
1 & 2	114	108.6	21.1	71	221
3 & 4	115	116.1	31.9	69	300
5 & 6	108	113.2	31.0	59	249

the test should have a binomial distribution with n = 63, p = 1/5. The pretest raw score distribution for first and second graders is shown in Figure 10. The histogram shown with dotted lines gives expected raw scores drawn as if, for example, one-sixth (or 19) of the children merely picked their answers at random. The average number of items gotten correct by guessing would be 13. Notice how many of the children did have pretest scores in the "guessing" range. Note that a raw score of 8 in a child of age 6 yields an IQ of 50, a raw score of 13 an IQ of 67, a raw score of 20 an IQ of 83.

Actually, it is rare that all children attempt all items. In this experiment, where teacher influences on subsequent test performance are of central importance, detailed data on test items answered incorrectly vs. items left unanswered at each testing should have been provided. It would be helpful in hypothesizing further about the nature of teacher effects, if found. Thorndike (1969) notes that the main influence of extra encouragement by the teacher might well be to increase the number of items attempted, even by guessing. RJ provide no data on this question, but Rosenthal notes elsewhere (1969, p. 690) that "... low IQs were earned because very few items were attempted by many of the children."

Reliability Questions

Examination of the score distributions reveals many extreme IQ scores less than 60 or greater than 160; RJ do not discuss these strange scores and have included them in standard analyses without comment. How stable are the IQ scores obtained across time? Test-retest correlations seem low at times especially for Reasoning IQ (see our table 6, RJ's table A-30). Looking at individual score sequences (using the data sent us by RJ) we noticed many instances of instability of IQ scores across time. A few examples of the more striking cases include one child with successive Total IQs of 55, 102, 95, 104, another with 84, 120, 107, 105, another with 88, 85, 128, 101 and another with 97, 88, 100, 127. For Verbal IQ we find sequences 54, 121, 101, 74 and 125, 87, 86, 68 and 167, 293, 174, 130. For Reasoning IQ, the sequences 0, 77, 82, 143 and 17, 148, 110, 112 and 111, 89, 208, 125 and 114, 81, 88, 106 appear. In view of the fact that children were tested three and four times with exactly the same test we should expect greater stability than

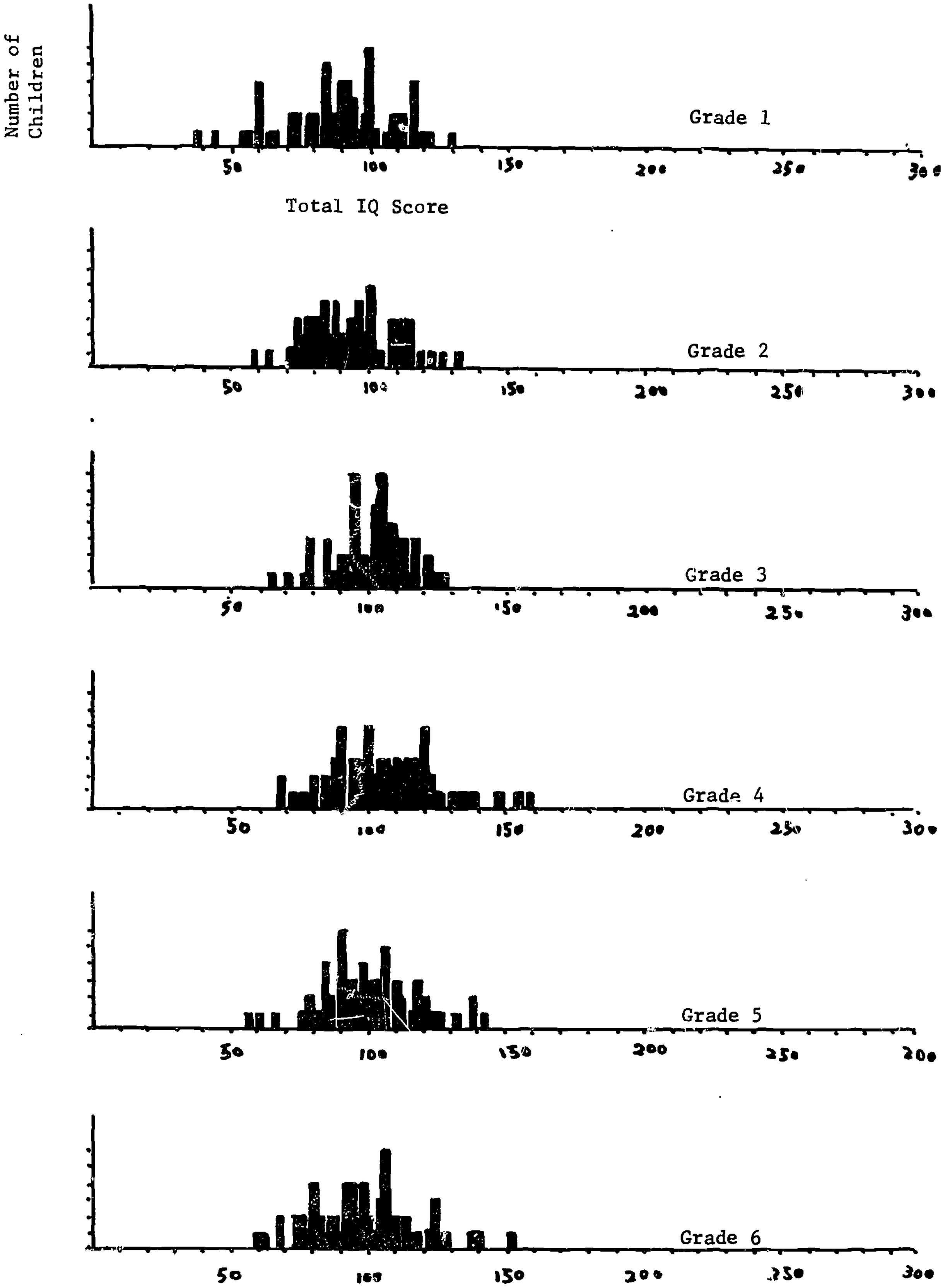


Figure 4: Pretest Total IQ Distribution by Grade

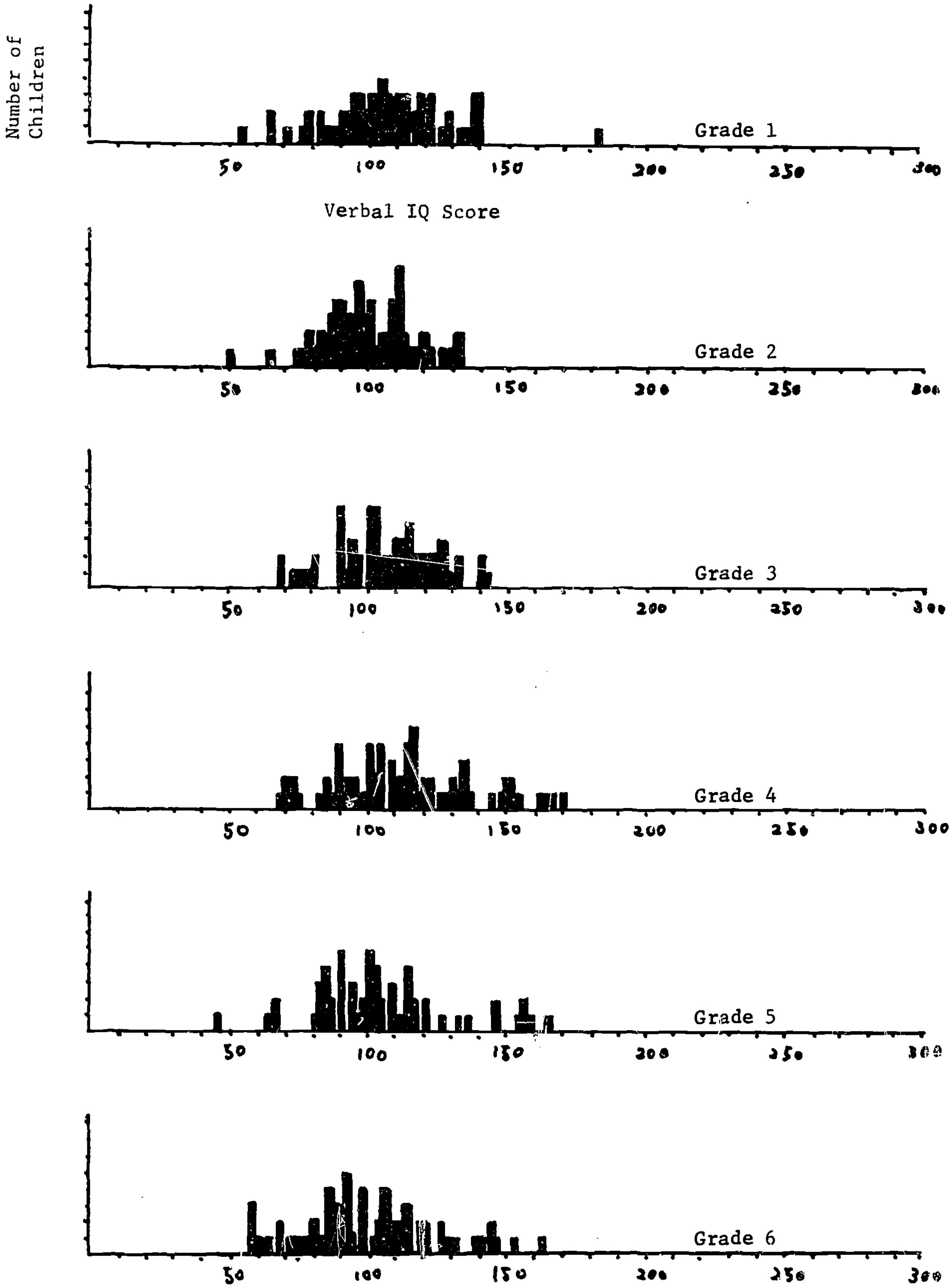


Figure 5: Pretest Verbal IQ Distribution by Grade

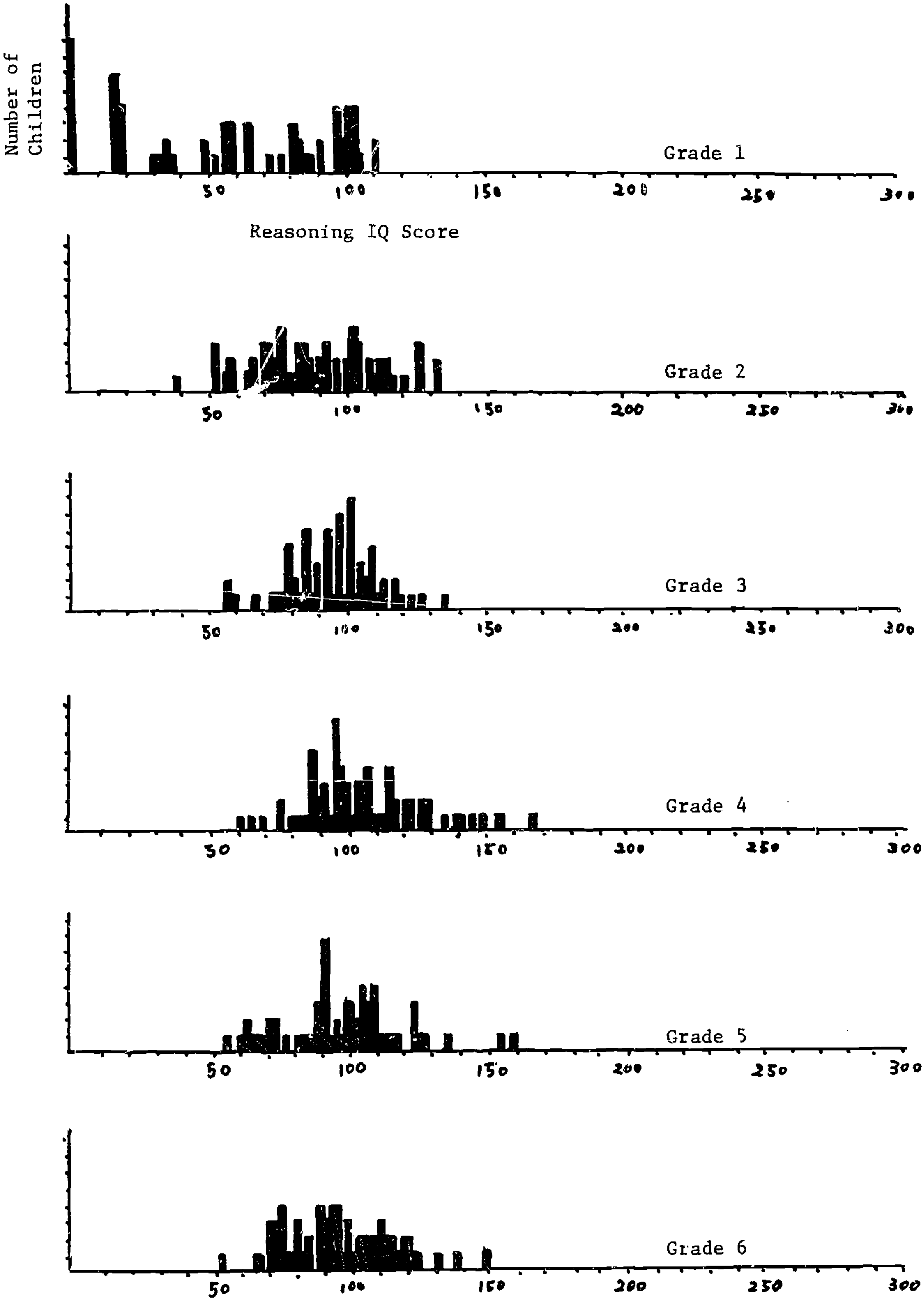


Figure 6: Pretest Reasoning IQ Distribution by Grade

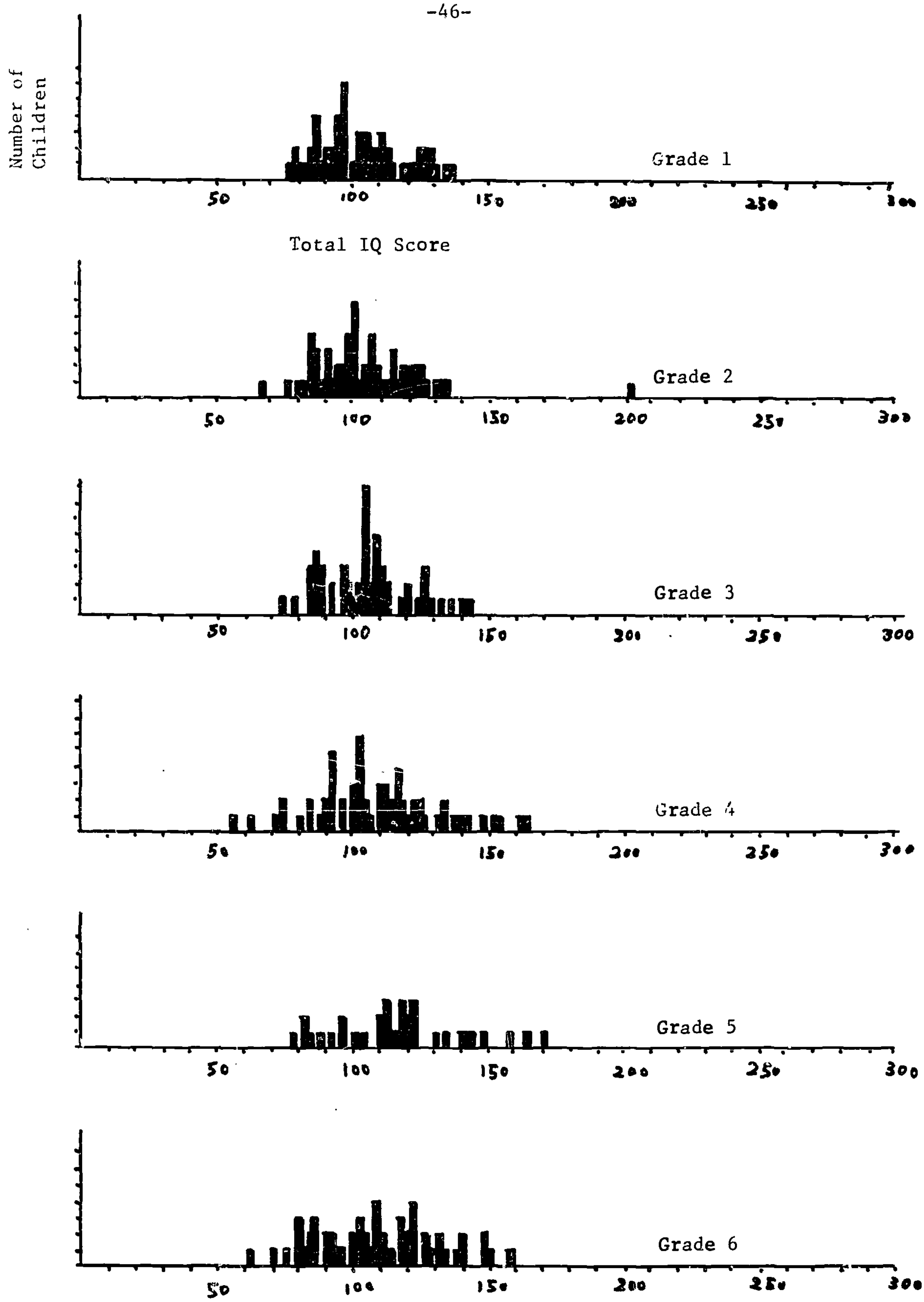


Figure 7: Posttest Total IQ Distribution by Grade

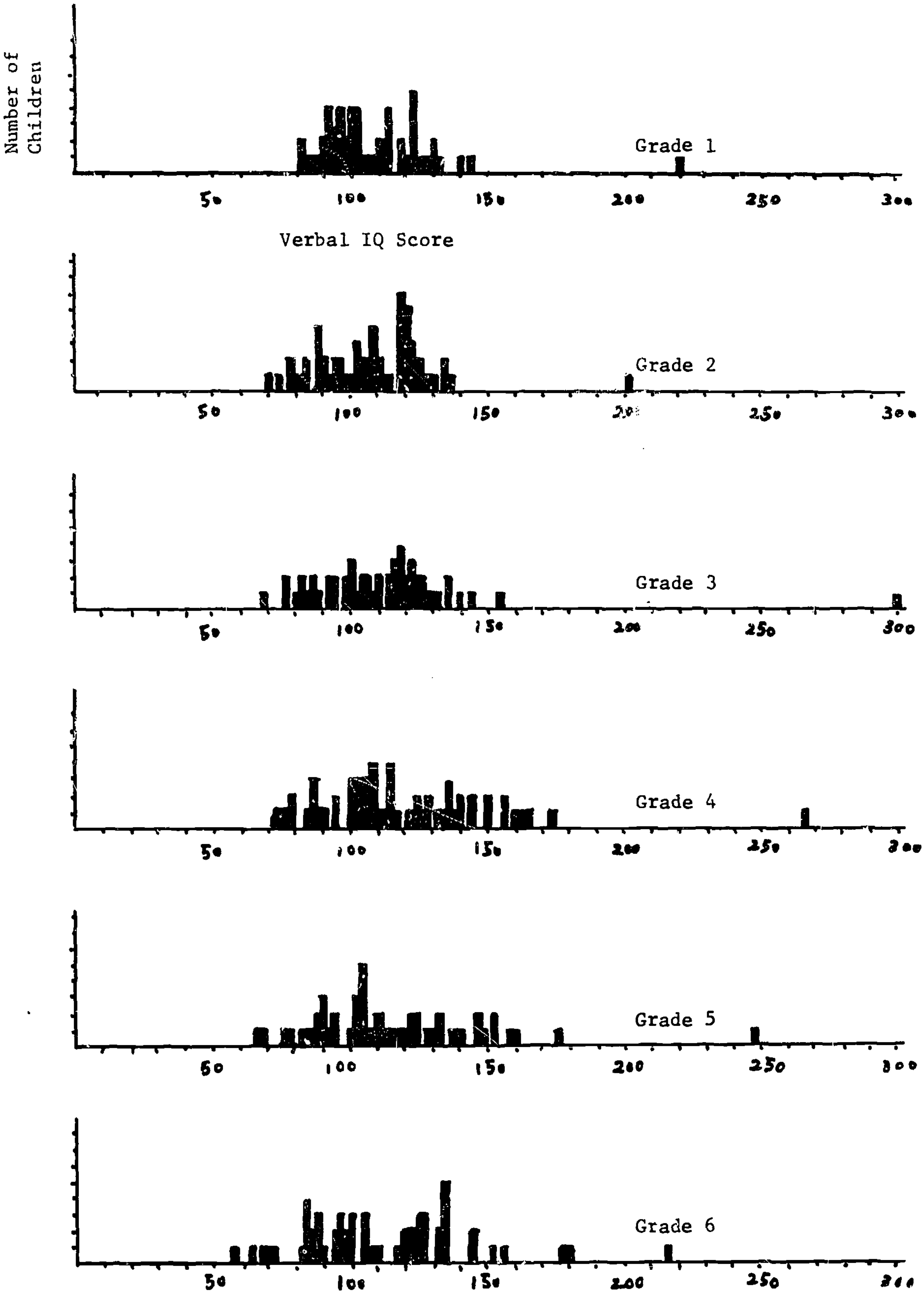


Figure 8: Posttest Verbal IQ Distribution by Grade

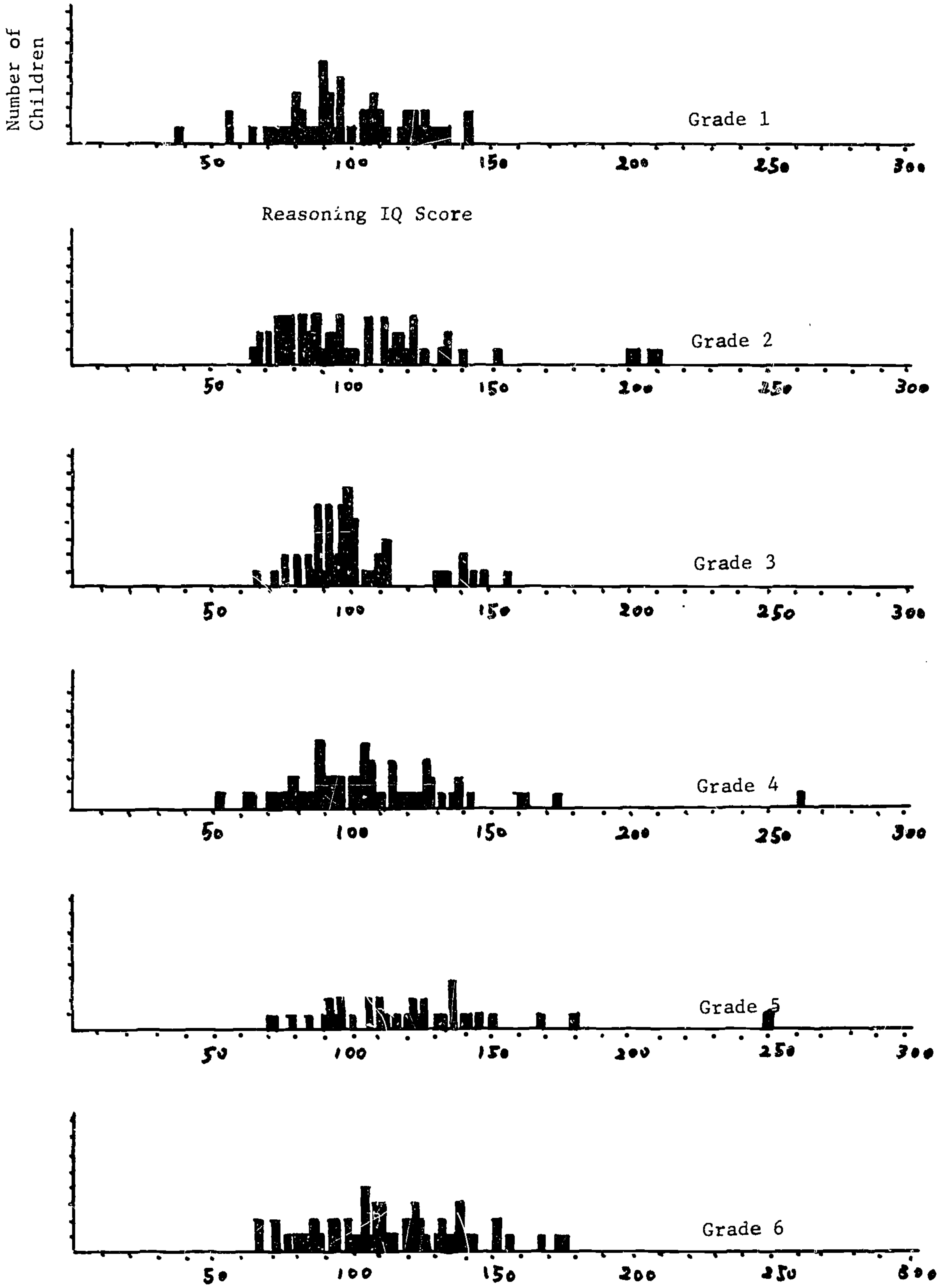


Figure 9: Posttest Reasoning IQ Distribution by Grade

this. A partial explanation of the unreliability of these scores is contained in the TOGA manual: "For second grade children of average or above-average ability, TOGA 2-4 will usually provide more reliable test scores."

The sections of the RJ book devoted to discussion of the reliability problem are unsatisfactory. RJ state:

> In fact, on a more rigorous basis, it can be shown that the less reliable a test, the more difficult it is to obtain systematic, significant differences between groups when such differences do, in fact, exist. In summary, there seems to be no way in which the 'unreliability' of our group measure of intelligence could account for our results although it could, in principle, account for the results not having been still more dramatic. (p. 149)†
>
> The problems of test unreliability ... were discussed and found wanting as explanations of our results. (p. 179)†

These statements are exaggerated and oversimplified. First, all statements about the effects of unreliability on a statistical test must be based on a probability model which describes the unreliability. The standard model for the reliability of gain scores is that pretest scores X and posttest scores Y come from a bivariate normal distribution with correlation coefficient ρ . (That is, X and Y both have normal distributions and are linearly related.) Thus "unreliability" is the same for all IQ levels, and the reliability, ρ , as well as the variances of X and Y, is the same for both experimental and control groups.

Under this standard model, it is true as RJ note, that the greater the unreliability of the test the larger the variance of gain scores and the larger the sample size necessary to show significance for true differences of a certain size between means of the groups. Therefore,

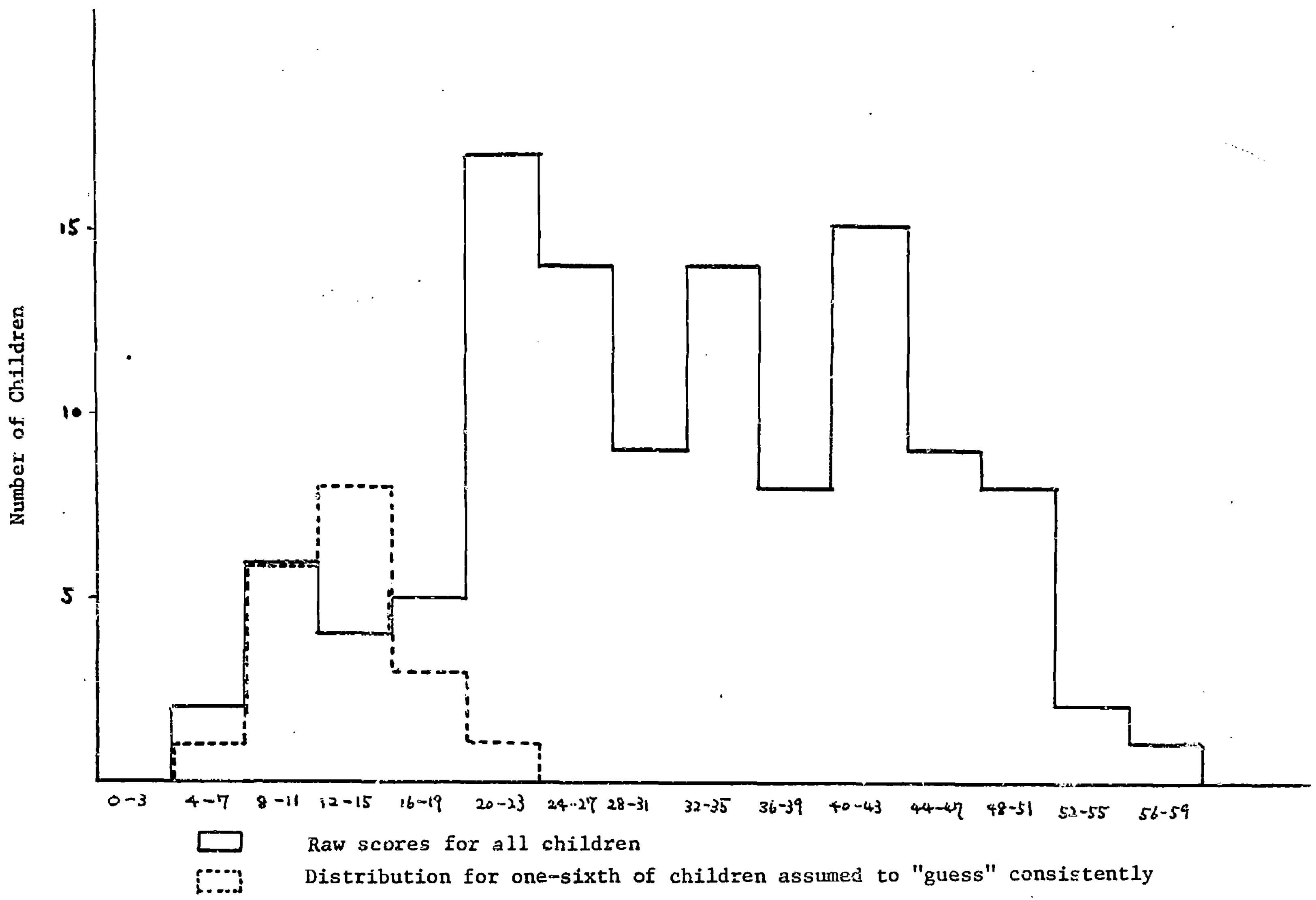


Figure 10: First and Second Grade Pretest Scores

Table 6

Test-retest Correlations

Pretest to Basic Posttest

	Control	Experimental
1st & 2nd Grades		
Total IQ	.66	.72
Verbal IQ	.73	.70
Reasoning IQ	.45	.50
3rd & 4th Grades		
Total IQ	.77	.87
Verbal IQ	.71	.74
Reasoning IQ	.57	.74
5th & 6th Grades		
Total IQ	.84	.87
Verbal IQ	.83	.85
Reasoning IQ	.63	.48

unreliability in a test increases the probability of Type II errors, that is, it increases the probability of finding no significant difference when true differences exist. However, it does not reduce the probability of a Type I error which is fixed by the experimenter; the probability of obtaining a statistically significant difference between experimental and control groups when no real difference exists is still equal to the p-value and is unaffected by the size of ρ . Furthermore, this is by no means the only possible model for unreliability and may not accurately describe the RJ data. The standard model maintains that IQ scores or gain scores for both control and experimental groups are drawn from the same distribution except that the means may be different. If the scores in the two groups come from distributions with different variances, different skewness, different kurtosis, then the actual probability of obtaining a significant difference in sample means when no difference in population means exists may be quite different from the nominal significance level of the test.

When two groups have markedly different sample sizes and markedly different variances, the actual significance level of a _t_-test may be quite different from the nominal significance level (see R. M. Elashoff, 1968). For example, if both the experimental and control groups have normal distributions with a ratio of sample sizes (n_c/n_e) of 5 and a ratio of variances (σ_c^2/σ_e^2) of .5, then in large samples and for a nominal significance level of .05, the actual significance level of the _t_-test would be .12. That is, to perform a _t_-test at the 5% level of significance, we reject the null hypothesis if the observed _t_-value is greater than 1.96. When $n_c/n_e = 5$ and $\sigma_c^2/\sigma_e^2 = .5$ the actual probability of observing a _t_ value greater than 1.96 under the null hypothesis is 12%. In the RJ experiment for the combined first and second grades, n_c/n_e is about 5 and the observed ratio of variances for Total IQ gain scores is $s_c^2/s_e^2 = .62$, consequently p-values quoted by RJ for comparisons in the lower grades are probably spuriously low.

Validity Questions

RJ do not provide a satisfactory discussion of the validity of their measure of "intellectual growth." "Intellectual growth" must mean more than changing a few answers the second time through a single test. Other mental ability information available from the school or obtainable without undue additional effort could have been used to examine the validity of the TOGA scores. A usual procedure in questions of construct validity is to show correlations between the measure in question and other indices presumed to represent the same or similar construct. RJ did not attempt to relate the TOGA scores to other acknowledged intelligence measures. The supporting evidence they introduce consists of changes in teacher grades, assessments of behavior made at one point in time, and a substudy of Iowa Tests of Basic Skills for the fifth and sixth grades. RJ report significant differences between experimental and control groups on one school subject out of eleven and three of nine "classroom behavior" indices. None of these differences, however, were as large as one point on scales of 1 to 4 for grades and 1 to 9 for behavior. No correlations between IQ and grades or behavior or achievement are shown; no correlations between gains in IQ and gains in grade points, changes in behavior, or gains in achievement are shown. In short, it is not clear how valid the TOGA IQ measures themselves are as a measure of intelligence or achievement or how valid changes in TOGA IQ scores are as a measure of intellectual growth.

In view of the conditions of test administration, pretest scores in the lower grades very likely involve variance due to differences in listening to instructions, perseverance, or resistance to distraction. These influences are particularly likely in the reasoning subtest, which is not teacher paced as the verbal subtest items are. Interpretations based on these influences would at least make the low pretest scores more credible, but a rather different interpretation of expectancy effects would also be required.

Rosenthal (1969) elsewhere argues that TOGA's validity is demonstrated by its correlation (.65) with ability track placement the following year. A test could predict a gross, three-level judgment of academic status well and still be nearly useless as a measure of individual intellectual ability or growth. Thus, such a correlation in no way validates the scale of measurement or its meaning and that is the question at issue here.

Another check on the relationship of the TOGA scores to other assessments of the children might be provided by considering track transfers. RJ do not discuss transfers of children between ability tracks, so the reader is permitted the dubious assumption that no students changed track across the study's two-year span even though some IQs changed more than 100 points. In fact, some track transfers did occur. According to information received from RJ the track location used in the analyses was track location as of January 1965, or about the time of the first posttest. There were indeed track changes during the experiment, however, as shown in Table 7. The relative numbers of control and experimental group children who changed tracks is consistent with their proportions in the experiment. Since the experimental group does not show a significantly greater proportion of upward changes than the control group, track changes do not support the contention that experimental children "benefitted more" than control children.

There is another difficulty created by the information that the track location is <u>not</u> that corresponding to the initial assignment of children within each class; we no longer know which class to compare these children with. Children have changed from cell to cell of the design during the experiment.

Another validity question concerns the experiment in general. In any experiment, one must be assured that the treatment conditions actually represent the intended variables. Particularly where incidental processes are of interest or where deception is involved, some procedure should be included to "cross-validate" the experimental effect. RJ took at least a first step in this direction by including a teacher interview and memory test at the end of the experiment. However, RJ fail to face the full implications of their results:

> While all teachers recalled glancing at their lists, most felt they paid little or no attention to them. Many teachers threw their lists away after glancing at them. (p. 154)†

Also, teachers could not recall with any degree of success which children had been expected to bloom and which had not.

> A memory test administered to the teachers showed that they could not recall accurately, nor even choose accurately from a larger list of names, the names of their own pupils designated as experimental-group children. (p. 69)†

Evidently the Pygmalion effect, if any, is an extremely subtle and elusive phenomenon that acts through teachers without conscious awareness on their part.

Table 7

Number of Children Changing Tracks During 1964-1965

	Control	Experimental	
No change	285	73	358
up	14	4	18
down	6	0	6
Total	305	77	382

CHAPTER V: REANALYSIS

In this section we discuss the methodological problems involved in the analysis of such a complex study, comment on RJ's choice of analysis, and present the results of our reanalyses.

The basic aim of analysis in the RJ experiment is to assess the relationship between pretest and posttest scores in the experimental and control groups, to locate any statistically significant differences between the groups, and to assess the practical importance of any significant differences observed. RJ based their analyses on the five-way classification of treatment x grade x track x sex x minority group status. They performed unweighted means analyses of variance using several different subsets of the classification factors because of unequal cell sizes and the prevalence of small or empty cells. The criterion was simple gain in IQ from pretest to posttest. Pretest to basic posttest ($T_3 - T_1$) is of primary interest but pretest to first posttest ($T_2 - T_1$) and pretest to follow-up posttest ($T_4 - T_1$) are included.

RJ have applied a standard analytic procedure, analysis of variance, without discussion of its assumptions or applicability and little attempt at exploration of the many other possibilities for analysis. Is an analysis of variance approach the most appropriate for this experiment? What about investigating the relationships between pre and posttest scores via regression analysis? What about analysis by classroom? What about nonparametric analyses?

Given the choice of a standard analysis of variance, we can ask whether these five particular factors should be included in the design. Can the number of cells be reduced in other ways than by dropping factors completely? Why choose simple gain scores as the criterion variable? Do the gain scores used satisfy the assumptions necessary for a standard analysis of variance to give valid results? Why not use posttest scores alone? covariance analysis? a repeated measures analysis? Is unweighted means analysis the appropriate way to calculate these analyses of variance: What about unweighted least squares? weighted least squares? While the main issue is whether analysis of variance is appropriate at all, we will also discuss the other questions.

Data analysis is an endeavor that must justify all that has preceded it in the experiment; analytic procedures must be chosen with the details of particular substantive hypotheses and the intricacies of appropriate statistical machinery clearly in mind. When considerable time and effort have been invested in the design and conduct of a study, hasty preplanned analysis is false economy at best and, at worst, risks gross misrepresentation of the data.

Most importantly, the researcher is not simply choosing a "test" to confirm some hypothesis. He is, or should be, investigating the heuristic value of alternative statistical representations of his data. As Tukey (1969, p. 90) notes:

> Data analysis needs to be both exploratory and confirmatory. In exploratory data analysis there can be no substitute for flexibility, for adapting what is calculated--and, we hope, plotted--both to the needs of the situation and the clues that the data have already provided. In this mode, data analysis is detective work--almost an ideal example of seeking what might be relevant.

Our reanalysis has two major objectives: 1) to provide a critical appraisal of the analytic approach taken by RJ and the conclusions warranted by the RJ data, 2) to discuss and illustrate the options available for exploring data of this type and the problems likely to be encountered with alternative approaches. As our discussion proceeds it will become clearer how crucial to the choice of analysis are the issues, raised earlier, of unbalanced sampling plan, 20% subject loss, and the measurement problems of extreme scores and unreliability.

In a complex unbalanced design with measurement problems, there is no one best way to analyze the data and the results may look rather different from one method of analysis to another. It would, in general, be preferable to analyze such data in several ways and compare the results. With imperfect data, potential problems associated with the application of particular methods may sometimes be balanced by comparing the results obtained from each. If the results are consistent across methods of analysis we can feel more secure about our conclusions. If not, the selection of which analysis is really most appropriate is crucial to the final conclusions. Choices must be made carefully and reasoning must be made explicit.

In this paper, we have reported the results of many different analyses and significance tests. They are included here to show the inconsistency of results from one method to another and are not necessarily valid analyses. That is, we cannot be sure how close the nominal p-value is to the actual probability of rejecting the null hypothesis when it is true. In fact, it is not clear that any analysis or significance test on these data can be accepted as wholly valid. It is only by examining the data from many different aspects that we are finally able to make any overall "conclusions."

The analysis section is organized as follows. First we suggest some procedures for handling the extreme scores. Second, we investigate the relationship between pretest and "basic" posttest scores for various subgroups and discuss the issues of choice of criterion variable and comparability of cells for an overall analysis. Then we report the results of using stepwise regression to estimate the size of the treatment effect. Our discussion of analysis of variance in unbalanced designs includes choice of factors and computation method and reports the results of some overall analyses and analyses within grade group. We also report an analysis using classroom as the experimental unit, and then offer a closer examination of the basic data for first and second grade children.

Extreme Scores

In the measurement section we noted the existence of many extreme scores in the RJ data. Very low scores are an indication that children responded randomly, consistently incorrectly, or did not respond at all to many questions; very high scores indicate that near the upper limits of the test the norming process is inadequate. Neither score gives an indication of the child's "true" mental ability. When there are so many extreme scores, it is difficult to know how to analyze the data. Even if we were to regard these scores as valid, their presence creates score distributions which are non-normal, skewed, and likely to have different variances in different subgroups. Applying standard statistical procedures to such scores may create a serious difference between the true and nominal significance levels of any statistical procedure (R. M. Elashoff, 1968). (See the section on reliability for an example of this.)

What procedures might be used to avoid such problems? Of course, the best way is to choose a measuring instrument and to plan data collection so that such scores do not arise. Perhaps the next best approach with the RJ data is to analyze the raw scores. This removes the problem of inadequate norming but forces us to analyze scores from the three different TOGA forms separately. As we shall see in later sections this is really necessary even using IQ scores. We have included analyses of total raw scores for first and second graders.

However, if analysis of the data in IQ form is still desired some procedure must be used to handle scores outside the main norming range of 60-160. One procedure is to truncate the data by excluding as too poorly measured any IQ scores outside this range. Another possibility is renorming the data by replacing all scores less than 60 by 60 and all scores higher than 160 by 160. Neither procedure is wholly adequate since the effect on various statistical approaches is unknown, but analyzing the data in all three ways, in original IQ form, in truncated IQ form, and in renormed IQ form provides information on the sensitivity of the results to the presence of extreme scores. Other possible procedures are trimming or winsorization, where a certain percentage of top and bottom scores are excluded or altered (see Dixon & Massey, 1969), and construction of a statistical model accounting for the presence of outliers (J. D. Elashoff, 1970).

Table 8 shows the effects of these three procedures on the test-retest correlation of total scores for first and second graders. Note that the values are highest using raw scores. Other differences in the effects of these options will appear in sections to follow.

Table 8

Test-retest Correlations for

First and Second Grades Total IQ

	Control	Experimental
Raw Scores	.73	.87
IQ Scores -- All	.66	.72
Renormed	.68	.75
Truncated	.70	.67

Relationships Between Pre and Post Scores

The basic aim of the RJ experiment was to assess the relationship between pretest and posttest scores in the experimental and control groups, to locate any significant differences between the groups, and to assess the importance of these differences. The first thing to do then is to examine the relationship between pretest and posttest in detail.

Regression Analyses. Scatterplots in Figures 11-19 show posttest IQ plotted against pretest IQ for Total, Verbal, and Reasoning scores for experimental and control groups of 1st and 2nd graders, 3rd and 4th graders, and 5th and 6th graders. This breakdown corresponds to the three different TOGA forms; further breakdown produces sample sizes too small for reasonable regression analyses. Experimental children are designated by X's, control children by dots. Norm limits are shown by the box drawn at 60 and 160 for both tests. The regression lines using all data and truncated data (all points outside the box deleted) are shown for both experimental and control children. Note that the lines labelled T are for truncated data. Figure 20 provides the scatterplot and regression analysis for total raw scores for 1st and 2nd graders.

Looking at the plots for first and second graders, one notices in Figure 11, for example, how strongly the one child with a posttest Total IQ of 202 affects the position of the regression line for the experimental group. The slope decreases from .93 to .58 when that one child is removed. The regression lines for experimental and control groups are generally closer together for the truncated data. Note that nearly 40% of the Reasoning IQ scores in Figure 13 appear well outside the norming ranges, most of them less than 60; 8 pretest scores are zero.

Is the relationship between pretest and posttest the same across treatments, grades, sexes? Are the relationships linear? Are the slopes near unity? How much do extreme scores affect the relationships? Tables 9 and 10 show regression slopes calculated using the original IQ data, renormed IQ scores, truncated IQ scores for each grade group, each treatment group, and Total, Verbal, and Reasoning IQs, as well as for some raw score data.

First, let us examine regression slopes for Total IQ in twelve groups--grade x sex x treatment, see Table 9. These twelve regression lines are significantly nonparallel, but within the six treatment by grade groups, there are no significant slope differences between the sexes. (Questions could be raised about the validity of the F tests for parallelism in view of the extreme scores; however, slopes for males and females seem generally close enough to warrant combining the sexes to obtain larger sample sizes.)

Accordingly, males and females were combined in subsequent analyses. With the sexes combined, we compared slopes for treatment and control groups. There was a significant difference in slopes only for the first and second grades (this difference is almost solely due to the one boy with a posttest IQ of 202), although the slope for the experimental group was slightly higher in all three grade groups. The major differences in slopes appear to be between grade levels, the slopes in the first two grades being considerably lower than those for the higher grades which are near 1.0. The same basic conclusions hold for Verbal and Reasoning IQ scores, although for Reasoning IQ the slopes are somewhat less than 1.0 even for the upper grades.

What effects do the extreme scores have on the regression slopes? Renorming and truncation procedures generally reduce the slopes and remove their apparent tendency to be higher in the experimental group. Except for the third and fourth grades, these procedures have reduced differences in slope between the experimental and control groups. Except for the first and second grade experimental group, different procedures produced very similar slopes for the reasonably reliable Total IQ but produced strikingly different slopes for Verbal and Reasoning IQ, which contained scores far outside the norming ranges. Examination of the scatterplots produces some doubt about assuming a linear relationship between pre and post scores for Verbal and Reasoning IQ.

Choice of criterion measure. To determine whether posttest scores for the experimental group are higher than for the control group, we must choose a grouping of the data (by classroom, by grade, etc.) and a criterion variable. We have a pretest measure T_1 and a posttest measure T_3. (The time 2 and time 4 IQ scores can be treated similarly. We ignore the repeated measures aspect of the data for the moment.) The three basic approaches are to examine T_3 (or posttest) alone, to use $T_3 - T_1$ (or simple gain), or to use T_3 with T_1 as a covariate.

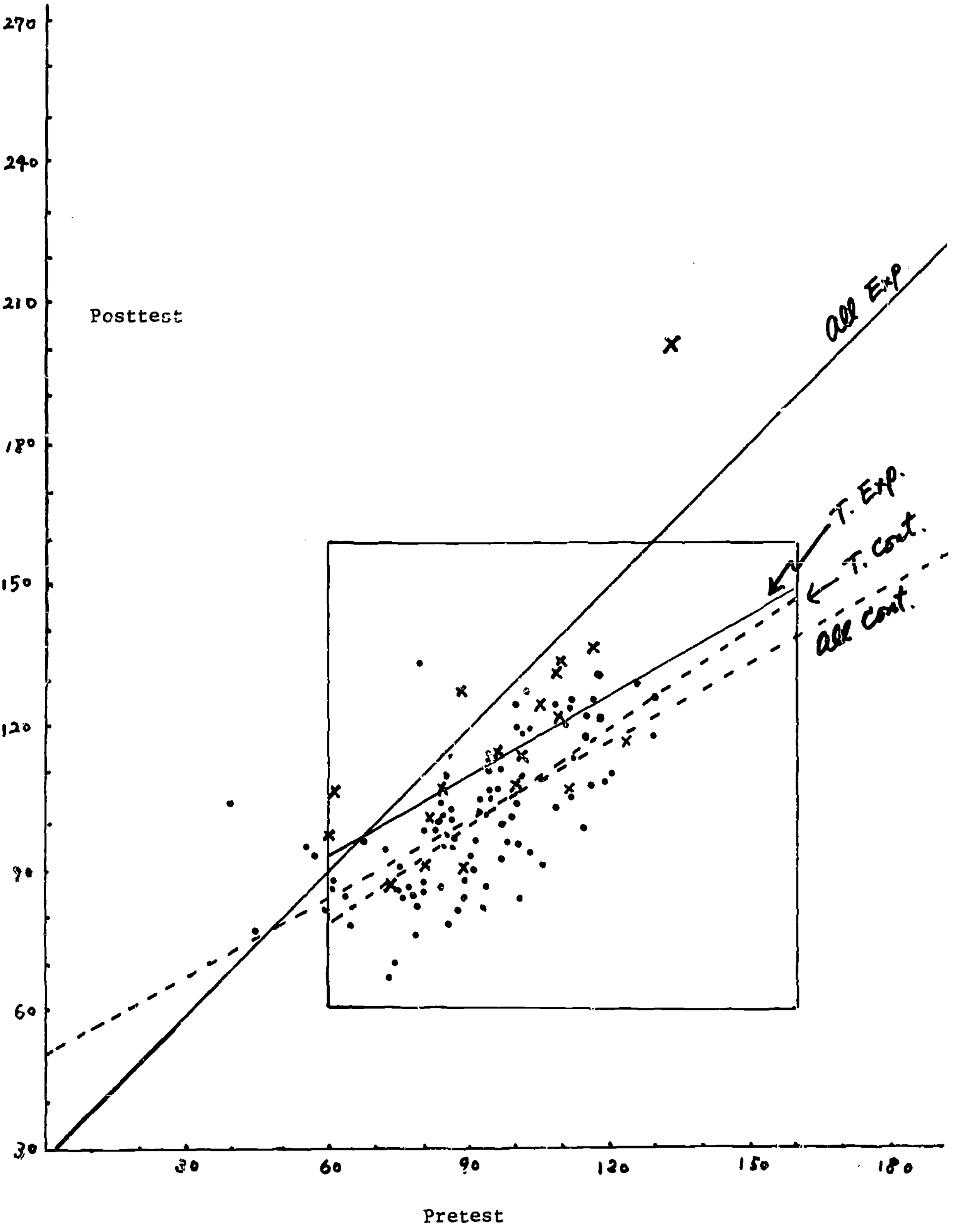


Figure 11: Total IQ Grades 1 & 2

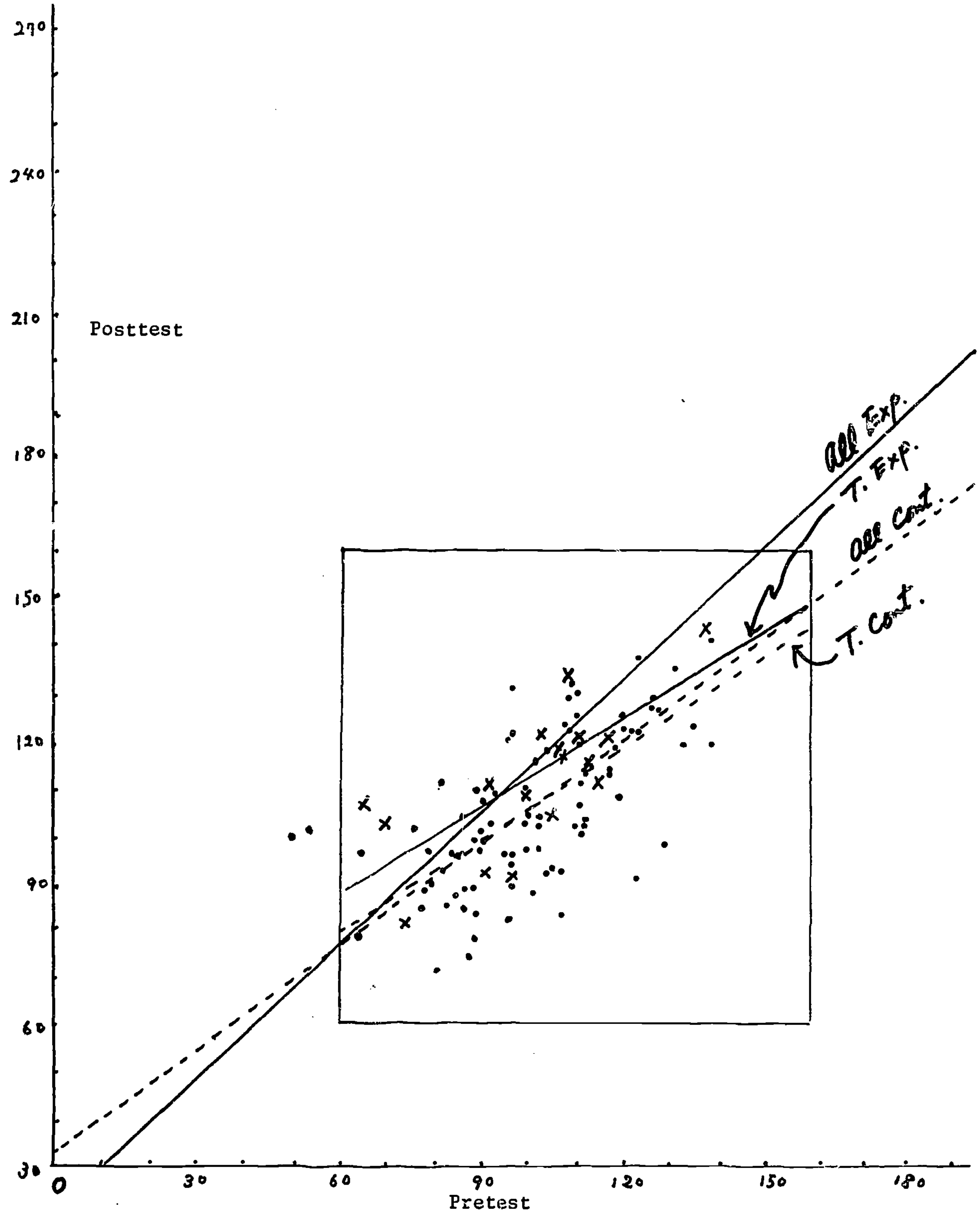


Figure 12: Verbal IQ Grades 1 & 2

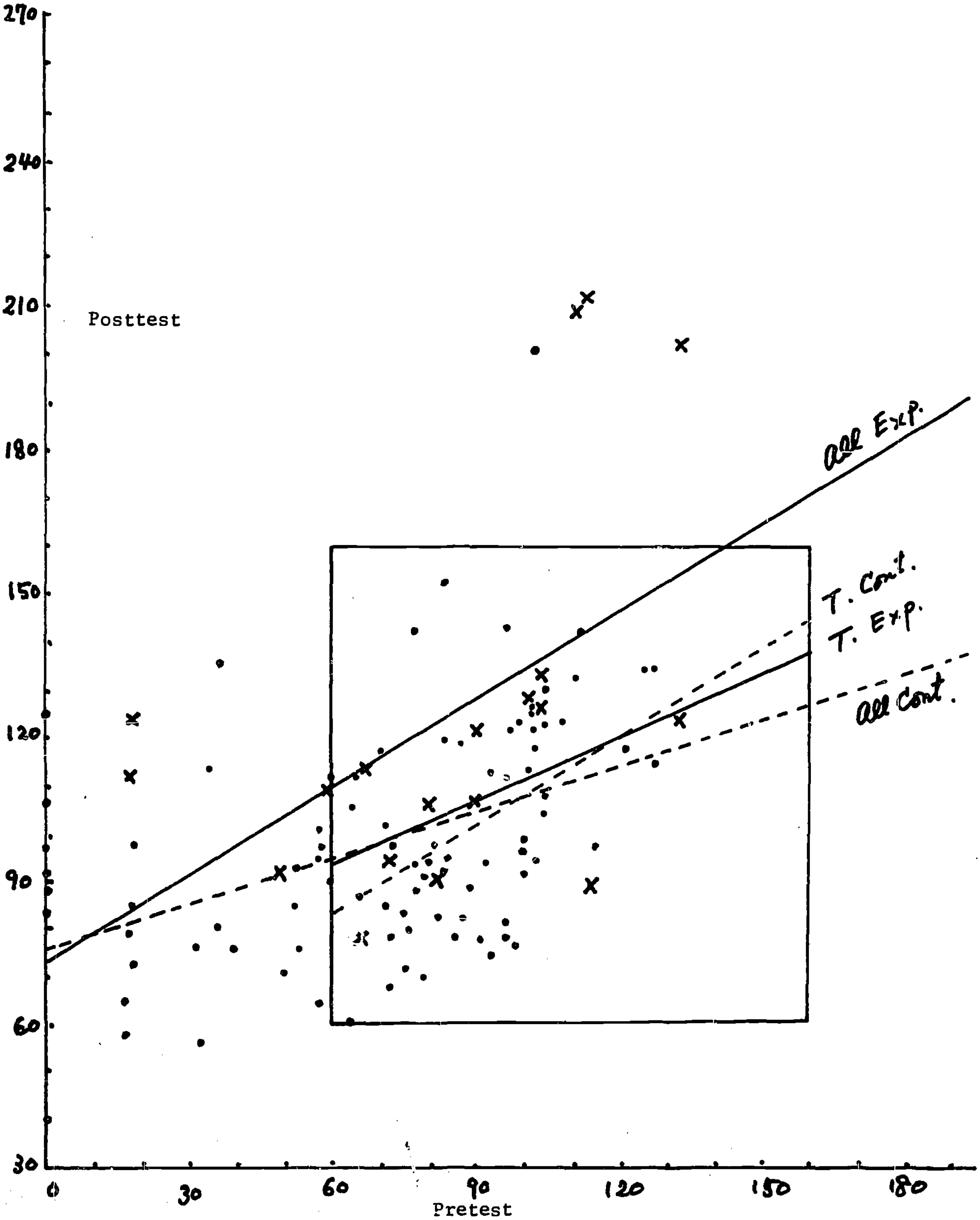


Figure 13: Reasoning IQ Grades 1 & 2

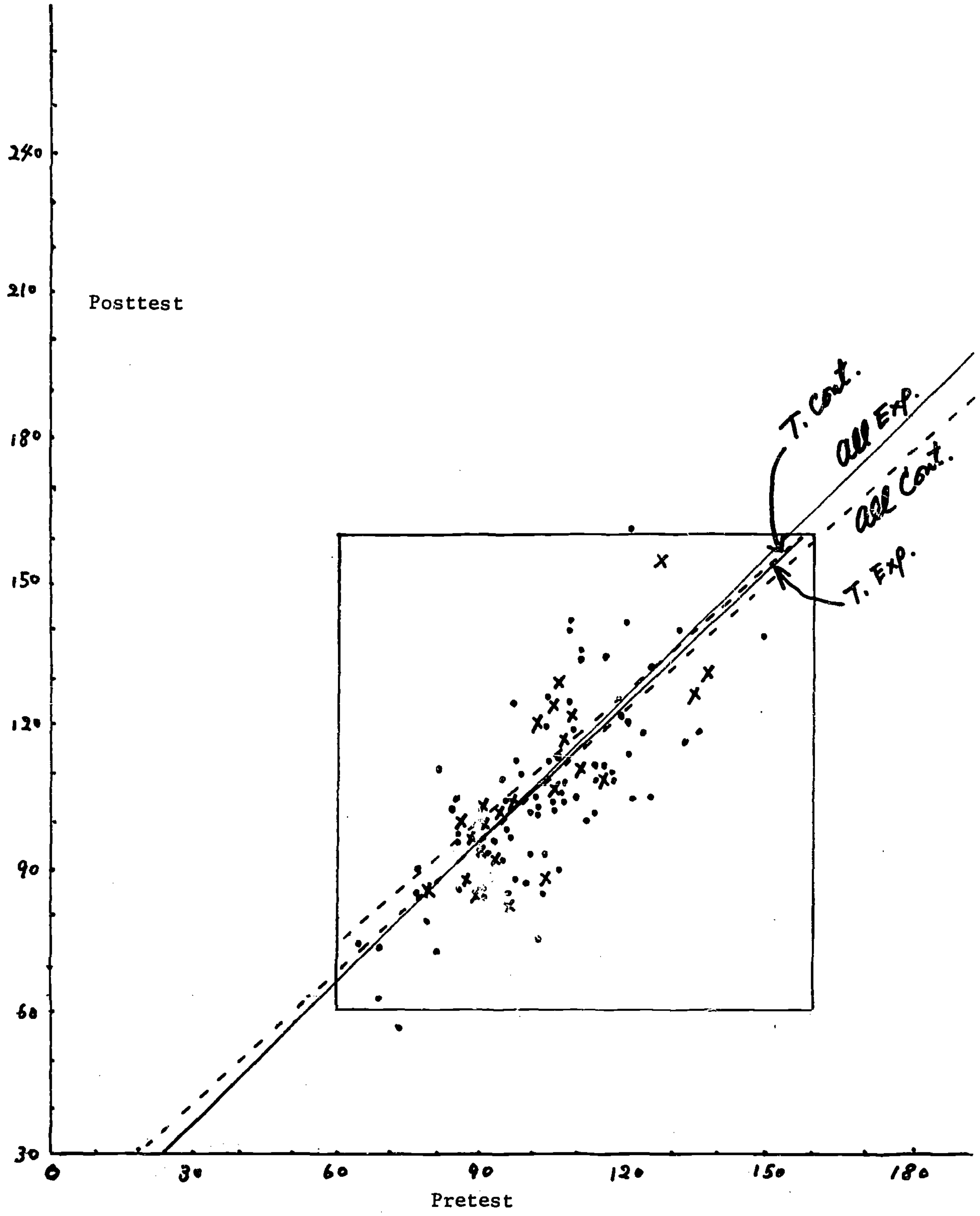


Figure 14: Total IQ Grades 3 & 4

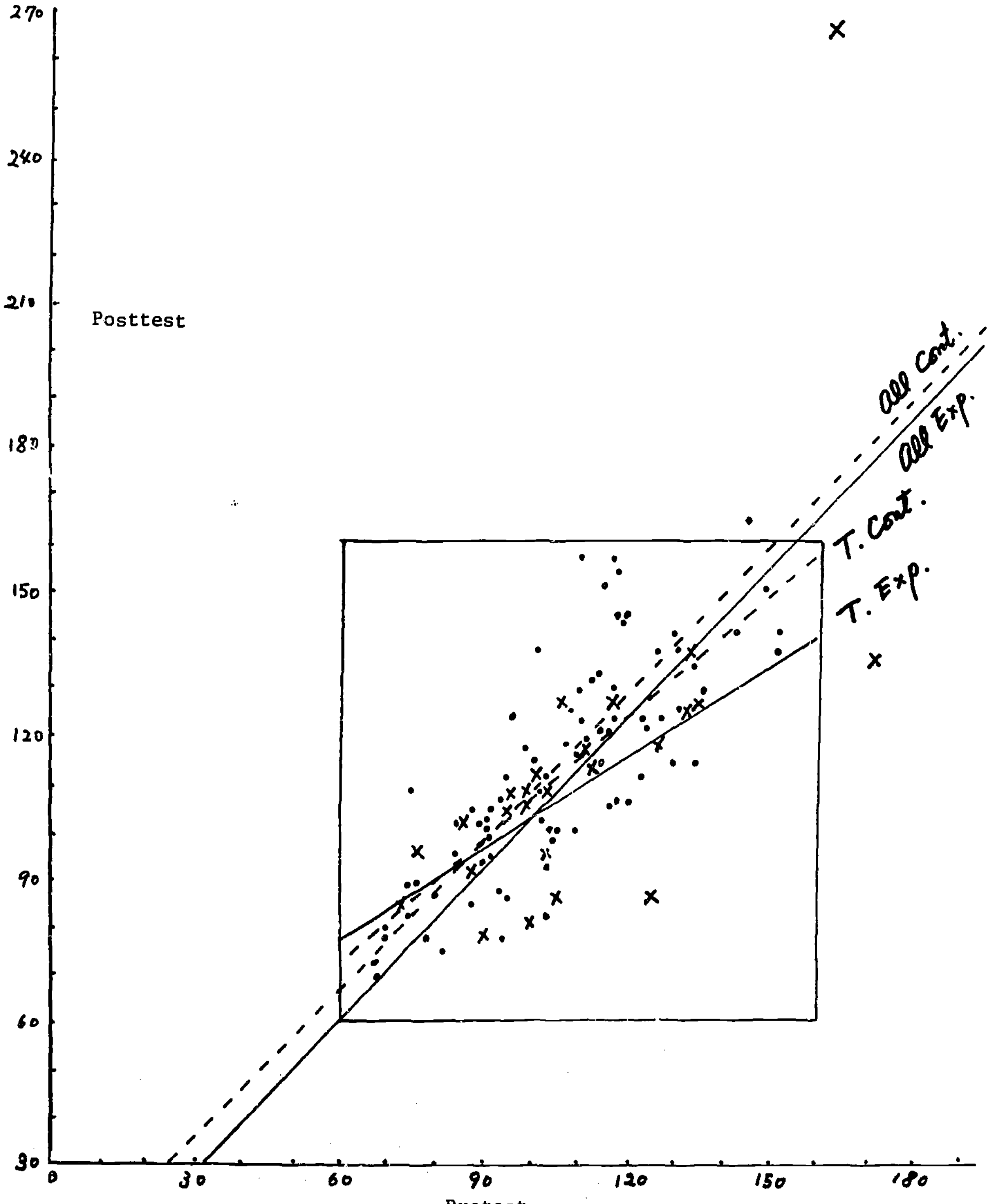


Figure 15: Verbal IQ Grades 3 & 4

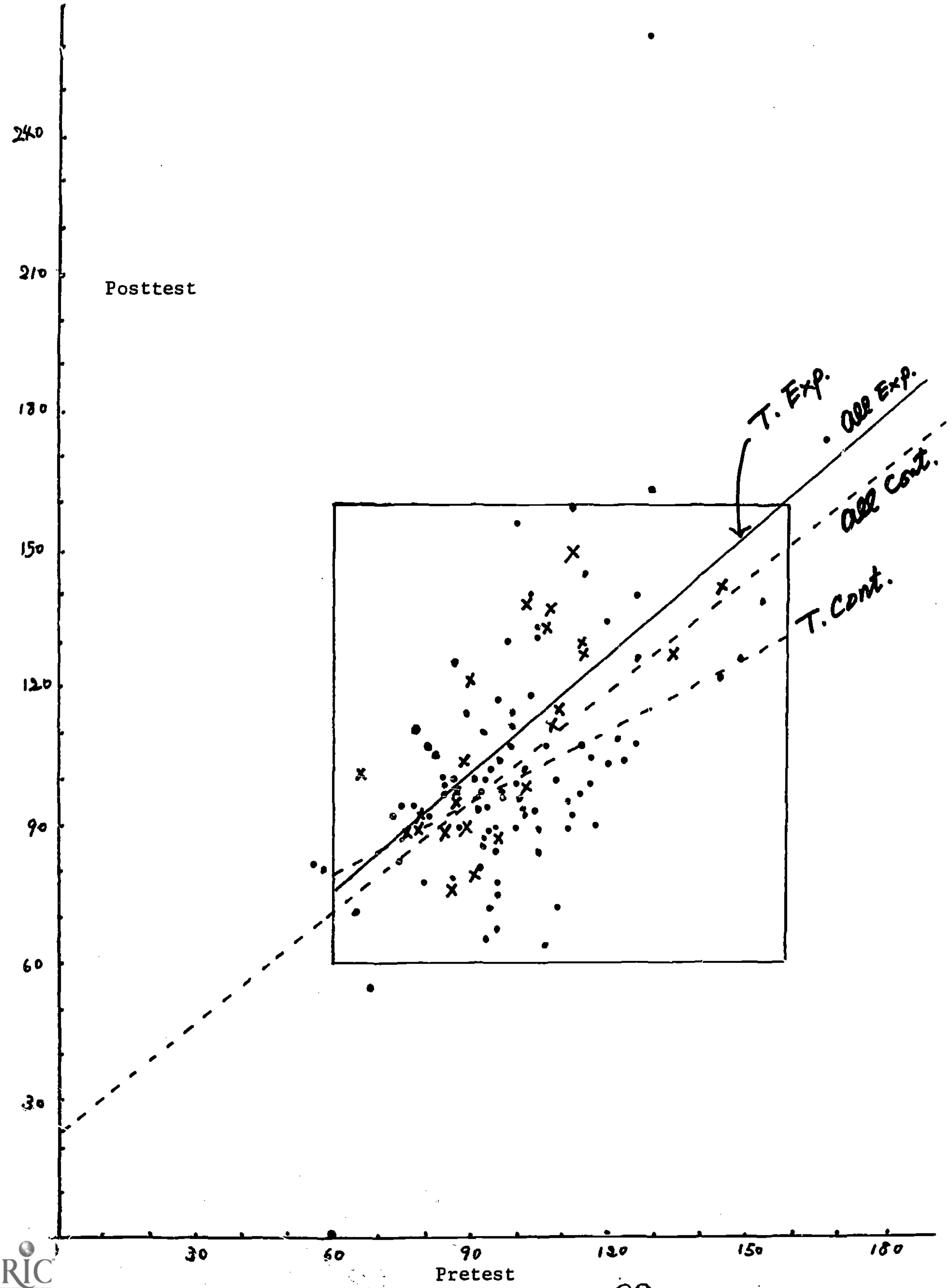


Figure 16: Reasoning IQ Grades 3 & 4

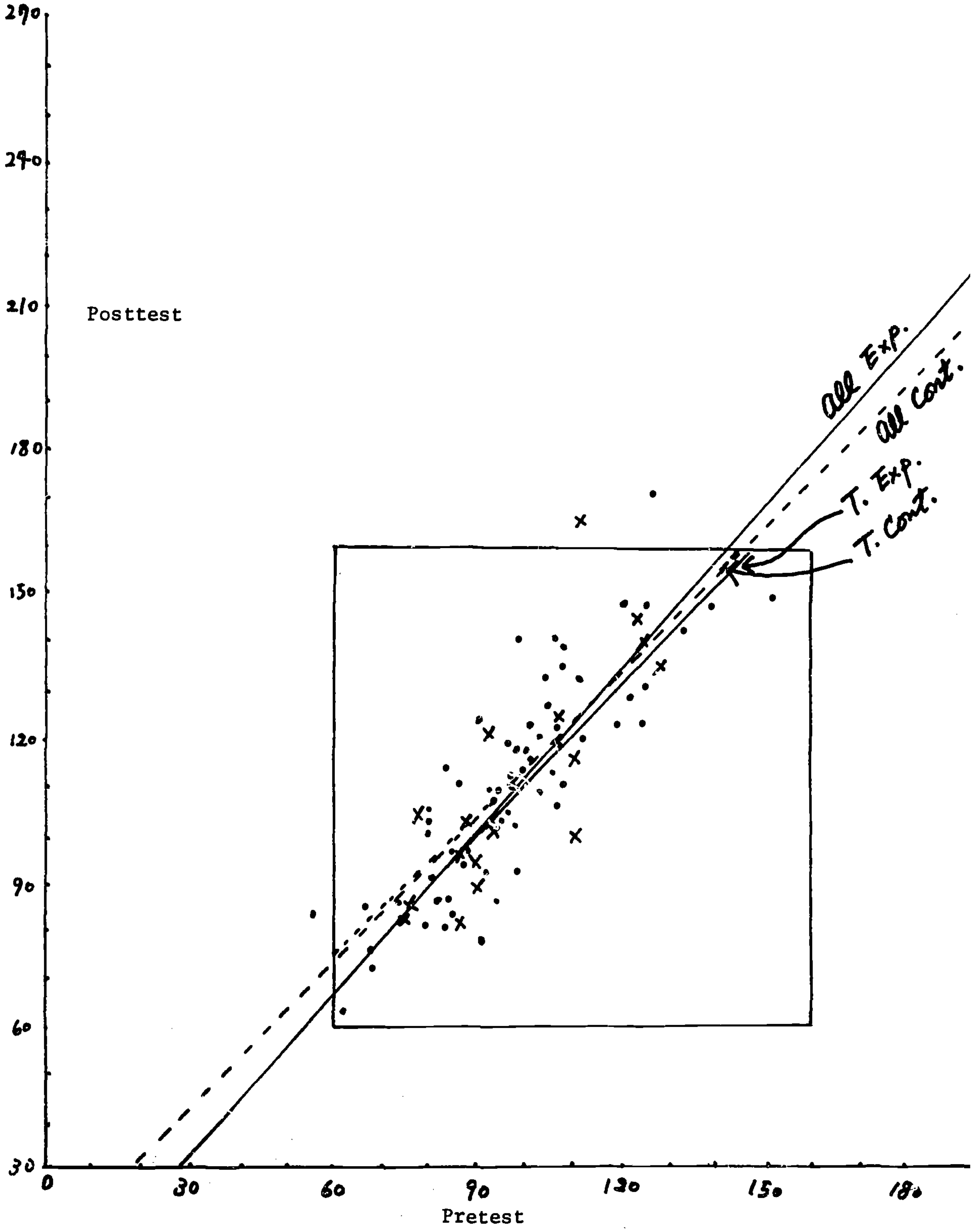


Figure 17: Total IQ Grades 5 & 6

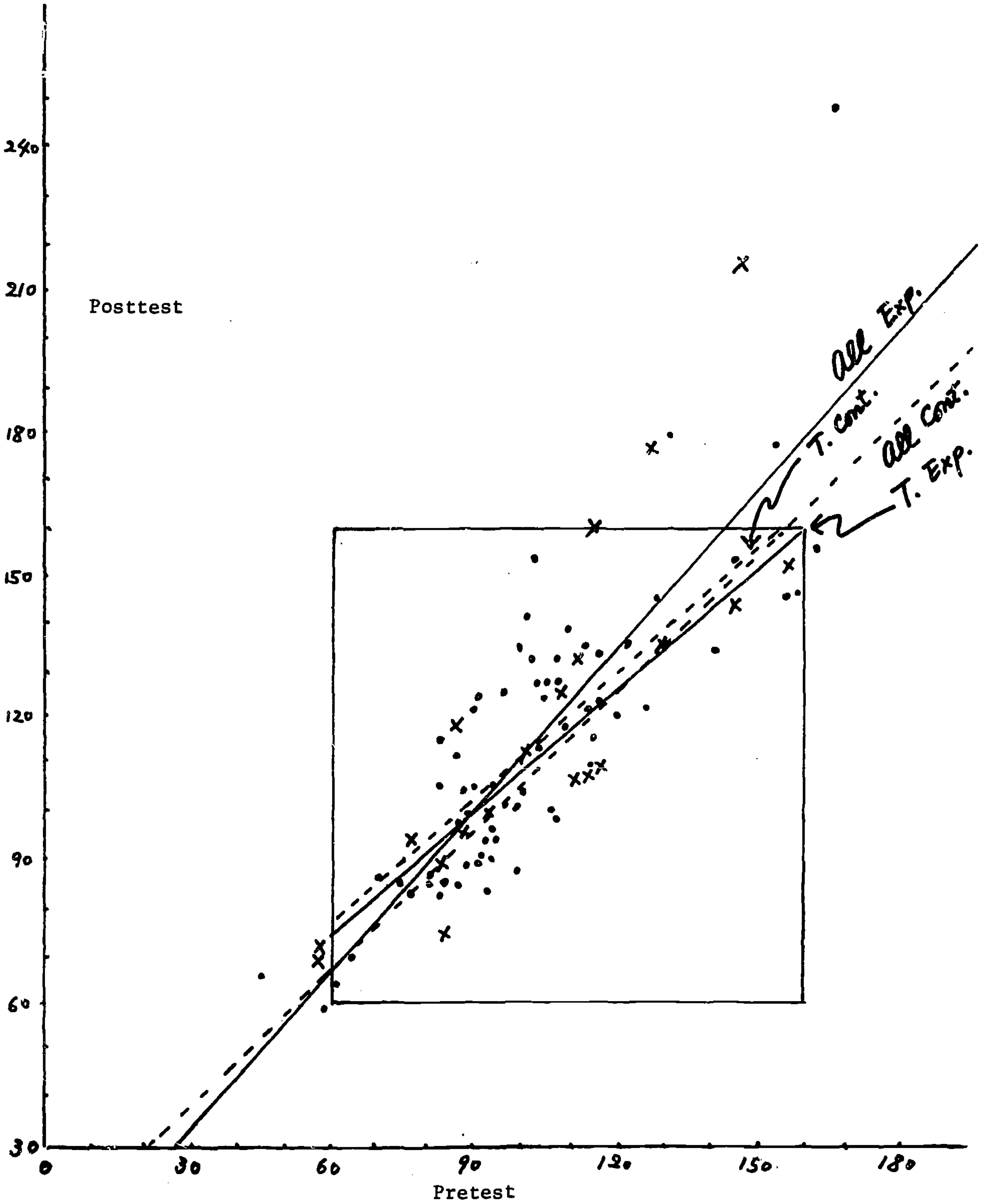


Figure 18: Verbal IQ Grades 5 & 6

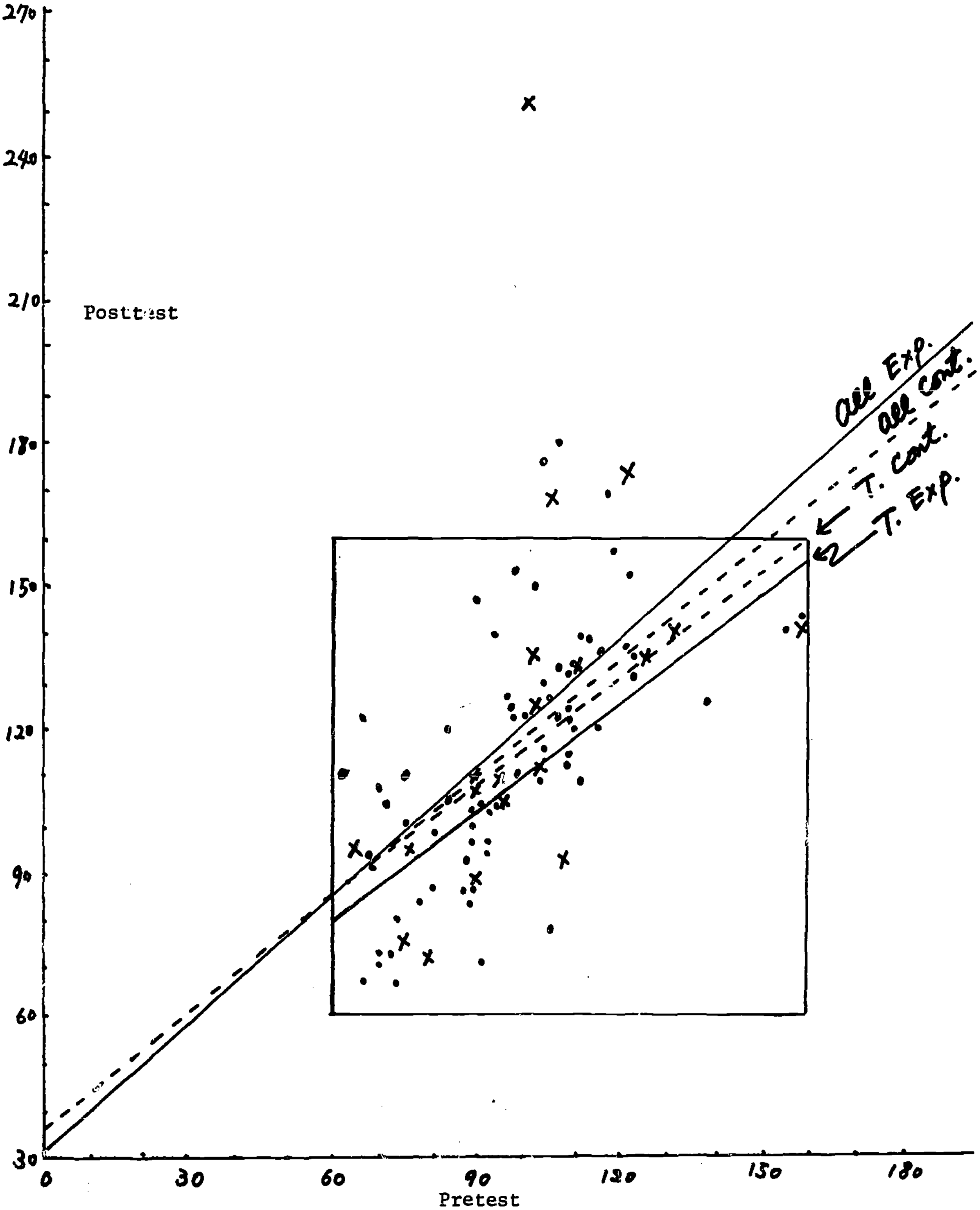
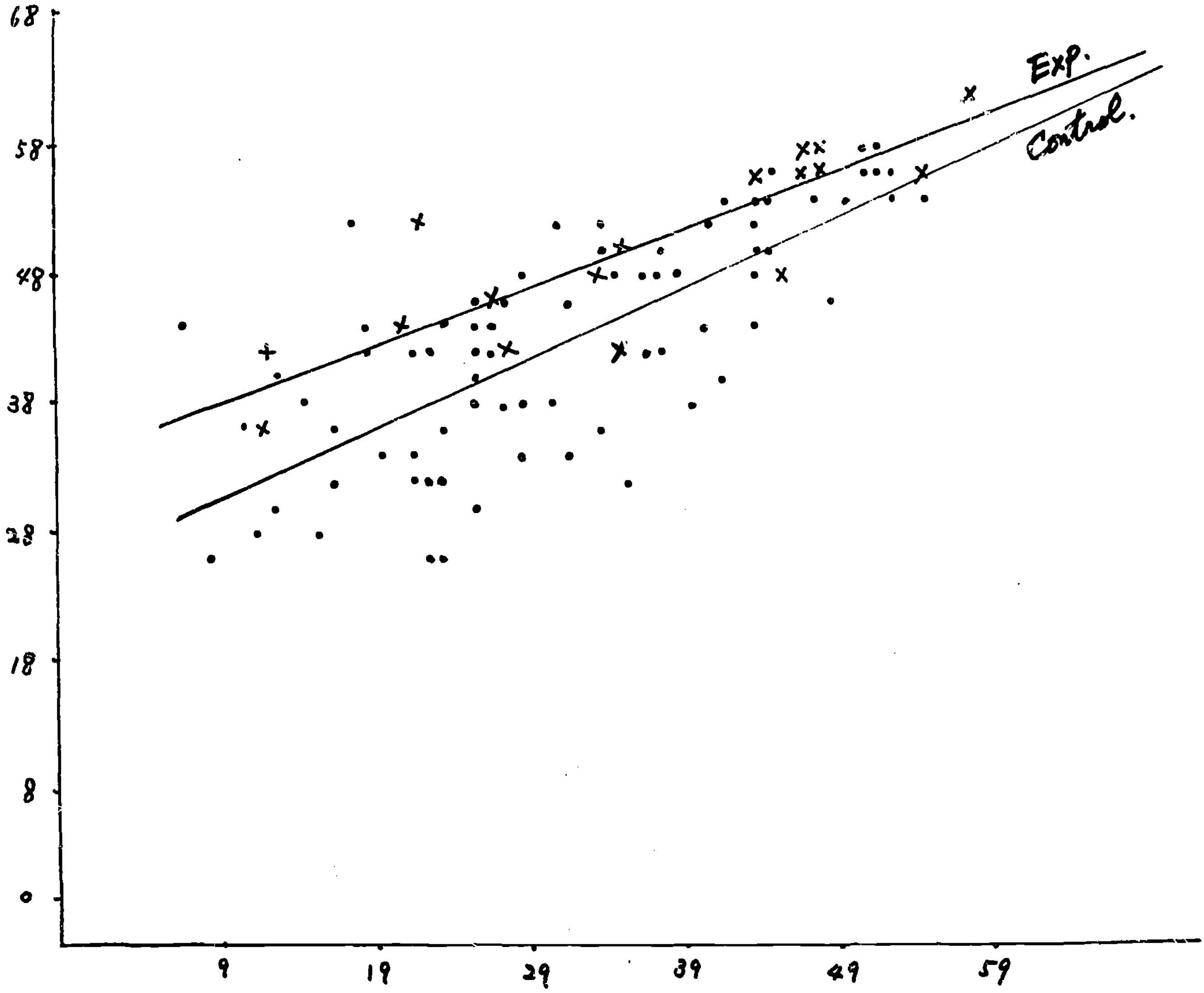


Figure 19: Reasoning IQ Grades 5 & 6

Posttest Raw Score



Pretest Raw Score

Figure 20: Total Raw Score First and Second Grades

TABLE 9

Slope of Regression Line for Sex by Treatment by Grade Group
Pretest to Basic Posttest

Total IQ

	Control		Experimental	
Grades	Female	Male	Female	Male
1 and 2	.62	.51	.72	1.03
3 and 4	.89	.92	1.12	.92
5 and 6	1.05	.94	1.14	1.07

These twelve slopes are significantly nonparallel $F_{11,296} = 2.59$ ($p<.05$)

TABLE 10

Slope of Regression Line for Treatment by Grade Group
Pretest to Basic Posttest

	Total IQ		Verbal IQ		Reasoning IQ	
	C	E	C	E	C	E
Grades 1 and 2						
IQ Scores	.56	.93	.72	.95	.32	.60
Renormed IQ	.62	.71	.63	.75	.58	.62
Truncated IQ	.69	.58	.66	.62	.61	.45
Raw Scores	.54	.45				
Grades 3 and 4						
IQ Scores	.90	.99	1.03	1.07	.88	.88
Renormed IQ	.89	.95	.91	.75	.71	.88
Truncated IQ	.84	.96	.87	.64	.53	.88
Grades 5 and 6						
IQ Scores	1.01	1.13	1.03	1.14	.82	.90
Renormed IQ	1.01	1.13	.90	.97	.81	.87
Truncated IQ	1.00	1.09	.87	.89	.76	.77

Each of these choices rests on an implicit set of assumptions about the data. If the particular assumptions necessary for an approach are not satisfied the results obtained by applying the approach may not be valid. We must examine the data to determine which approach is most appropriate.

RJ rely solely on simple gain scores $T_3 - T_1$ arguing that "... posttest only measures are less precise than the change or gain scores...." (p. 108)† As we shall see this oversimplified claim is actually false for the Reasoning IQ scores.

Using posttest only (T_3) as a criterion requires the fewest assumptions. Assignment to treatment must be random and score distributions should be approximately normal with similar variances in both groups. We note that where the sample sizes of the two groups are quite different, as in the RJ study, this assumption of equal variances is much more important. Potentially, analysis of variance of T_3 only is the procedure most seriously affected by initial differences between groups. For comparison with other methods assume that the within-group variance using posttest scores is σ_ε^2 .

If the within-group correlation between pre and posttest scores, ρ , is high, gain scores and covariance analysis can be expected to be more precise than analysis of variance of posttest scores. Using either gains or covariance requires random assignment to treatments and a similar relationship between pre and post scores in both groups. To derive formulas for the precision of gain scores or covariance analysis, we must adopt a model for the relationship between pre and posttest scores. We follow the general formulation of Cochran (1968) and assume that in the absence of measurement errors, y or posttest has a linear regression on x (pretest)

$$y = \alpha + \beta x + \varepsilon \ .$$

The observed scores, X and Y however, do contain measurement error

$$Y = y + u$$

$$X = x + v$$

and we can write:

$$Y = \alpha' + \beta' X + \varepsilon' \ .$$

Under certain general conditions of independence and normality of variables, we find that the residual within-group error variance in covariance analysis will be about

$$\sigma^2_{\varepsilon'} = \sigma^2_{\varepsilon} (1 - \rho^2 R_X R_Y)$$

where ρ is the correlation between y and x and R_X and R_Y are the reliabilities $(R_X = \frac{\sigma^2_x}{\sigma^2_x + \sigma^2_v})$. (Note that the correlation between observed scores X and Y is $\rho\sqrt{R_X R_Y}$.)

Use of covariance analysis rests on a number of important assumptions about the underlying structure of the data (J. D. Elashoff, 1969). In the absence of measurement error $(R_X = R_Y = 1)$, then, covariance analysis can be expected to reduce the error variance by about $100\rho^2$%; thus ρ must be larger than .3 for covariance analysis to reduce the error variance appreciably. The less reliable the pretest and posttest the greater ρ must be before covariance will be much more precise than analysis of variance on posttest scores alone; in addition;

when the pretest is measured with error, covariance procedures generally underestimate the slope and undercorrect for pretest differences.

The use of gain scores makes the implicit assumption that $\beta' = 1.0$, i.e., that the regression of observed posttest on observed pretest has a regression slope of unity. If this is the case, analysis of variance of gain scores will give nearly the same results as analysis of covariance. If not, the error variance can be expected to be about

$$\sigma_g^2 = \sigma_\varepsilon^2 \{\frac{(2\beta'-1)}{\beta'^2} (1-\rho^2{}_{R_Y R_X}) + \frac{(\beta'-1)^2}{\beta'^2} \}$$

which is always greater than $\sigma^2_{\varepsilon'}$ for $\beta' \neq 1$. Note that these variance figures are derived for large samples; for smaller samples imprecision due to the estimation of β will make $\sigma^2_{\varepsilon'}$ larger. Little is known about the comparative robustness of these two procedures. Comparisons of two groups using gain scores will be misleading when the regression slope of post on pre is not unity for both groups or the pretest score distributions are different in the two groups; since in either case their use would not properly adjust for pretest differences. In a general discussion of this topic, Cronbach and Furby (1970) have suggested that gain scores are rarely useful for any purpose in educational research.

Using these formulas, we can predict whether posttest scores or gain scores will have smaller error variance for the RJ experiment by referring to evidence contained in RJ's Table A-30. We find a pretest-posttest correlation for the total school of approximately .75

for Total IQ and Verbal IQ but only about .50 for Reasoning IQ. Thus assuming that $\beta' = 1$, using gain scores should provide a decrease in error variance of about 50% for Total IQ and Verbal IQ and none at all for Reasoning IQ. Referring to Table 19 in the analysis of variance section (our page 101), we find that for two types of analysis of variance actually performed the decrease in error variance obtained by using gain scores was about 33% for Total IQ and 50% for Verbal IQ but that error variance increased by about 8% for Reasoning IQ. So, for Reasoning IQ, a posttest criterion is not less precise than a gain criterion. (Differences between the predicted and observed decreases in error variance occur because the formulas are for large samples, and because the correlations taken from Table A-30 were computed with all groups combined while the correlation in the formula is the within group correlation.)

Thus, careful examination of these score distributions, scatterplots, and regression slopes suggests which scores are reasonable to analyze, whether grades (or TOGA forms) can be combined, and which analytic procedures seem appropriate.

If IQ scores are to be used, all analyses should be based on Total IQ; Verbal and Reasoning subscores are unreliable and inadequately normed in all grades. The only overall analysis combining all grade groups that seems reasonably justified is analysis of posttest Total IQ scores. If random assignment to treatments can be assumed, analysis of posttest Total IQ scores is unbiased. In view of the lack of assurance on this question, however, and the higher pretest scores shown by the experimental group (see Tables 20-22), the results of such an analysis must also be interpreted with caution. Covariance analysis or gain score analysis using all grades is unwise because of the dissimilarity in pre-posttest relationships across grades. Using raw scores, the three forms of TOGA are not comparable.

Grades 3 and 4 and Grades 5 and 6 might reasonably be combined and analysis of Total IQ here, using covariance analysis, (or analysis of variance of gains) would not be unreasonable. There seems little reason to perform separate analyses for males and females. Grades 1 and 2 present a more difficult problem, however. Here, gain scores are especially suspect because the pre to posttest slope is substantially less than one and the groups differ on the pretest. Covariance analysis should not be used with all IQ scores included because of the difference in slopes between groups, though it might be useful for renormed or truncated scores. Both posttest only and covariance analysis may be inadequate because of the large group differences in the pretest, as well as its unreliability. Analysis using raw scores seems most desirable. This could eliminate some of the problems caused by inadequate norming of the test. Test-retest correlations are higher for raw data and the regression slopes between pre and posttest are similar for experimental and control groups.

Investigation of Treatment Effects Using Stepwise Regression

It is most importunt to assess the magnitude of any "significant" treatment effects observed. One approach to this problem is stepwise regression, see Appendix A. Taking posttest IQ as the dependent variable, we can determine how much of the variance in posttest scores is accounted for by linear regression on pretest IQ scores, treatment, sex, and other interesting variables.

First, we performed separate analyses for each of the three grade groups using the third or "basic" Total IQ score as criterion. Pretest Total IQ, treatment group, track, sex and minority-group status were included as predictor variables. In the analysis, pretest Total IQ was forced into the equation first and treatment was second; the other variables were left free to enter in any order. Results are shown in Table 11. These analyses must be interpreted with caution because of the extreme scores in Total IQ for grades 1 and 2 and because the other variables are categorical. In addition, for a dichotomous variable such as treatment, R^2 is lower when the number in each group is not the same than when the split is 50-50; R^2 for a 20-80 split will be roughly 2/3 of R^2 for a 50-50 split given the same difference in Total IQ means. In addition the predictor variables are not independent and their contributions overlap. Thus these analyses must be regarded as giving at most a rough approximation of the relative importance of the predictor variables. Pretest Total IQ predicts 43%, 63%, and 72% of the variance in posttest Total IQ for grades 1-2, grades 3-4, grades 5-6, respectively. Including all the variables accounts for a total of 55%, 70%, and 75% respectively of the variance in posttest. For grades 3-4 and 5-6, treatment accounted for less than 1% of the variance in posttest Total IQ scores; treatment accounted for 7% of the variance in grades 1-2. No attempt has been made to assess the statistical significance of these increases in R^2 because of the difficulties mentioned earlier. Our only purpose is to gain an impression of the relative importance of any treatment effect.

As we remarked earlier, total raw scores seemed a more desirable criterion measure than Total IQ for grades 1 and 2. The same type of

TABLE 11

Results of Stepwise Regression Analyses for Grade Groups 1 and 2, 3 and 4, 5 and 6

Criterion Variable: Total IQ on Basic Posttest

Predictors: Total IQ Pretest, Treatment, Track, Sex, Minority-Group Status

Criterion	Step	Variable Entered	F to enter	R^2	Increase in R^2
Grades 1 & 2	1 forced	Total IQ1	85	.43	.43
Total IQ 3	2 forced	Treatment	15	.50	.07
	3-5 free	sex, track, minority		.55	.05
Grades 3 & 4	1 forced	Total IQ1	190	.63	.63
Total IQ 3	2 forced	Treatment	.5	.63	.00
	3-5 free	track, sex, minority		.70	.07
Grades 5 & 6	1 forced	Total IQ1	226	.72	.72
Total IQ 3	2 forced	Treatment	.0	.72	.00
	3-5 free	track, minority, sex		.75	.03

analysis was repeated for grades 1 and 2 using total raw scores with age and grade included (Table 13). All variables were forced to enter in the order shown; treatment was entered third in the first regression and was forced to enter last in the second regression. Note that using raw scores, the pretest predicts 55% of the variance in posttest and all variables together predict 65% of the variance. The partial correlation

TABLE 12

Results of Stepwise Regression Analyses for Grades One and Two

Criterion Variable: Total Raw Score on the Basic Posttest

Predictors: Pretest Raw Score, Treatment, Track, Sex, Minority-Group Status, Grade, Age

Criterion	Step	Variable entered	F to enter	R^2	Increase in R^2
Total Raw Score on Basic Posttest	1 forced	Pretest raw score	136	.549	
	2 forced	Age	0	.549	.000
	3 forced	Treatment	9.3	.584	.035
	4-7 free	sex, track, minority, grade		.654	.070
	1 forced	Pretest raw score	136	.549	
	2-6 forced	Age, grade, etc.		.617	.068
	7 forced	Treatment	11.2	.654	.037

of age with posttest after pretest has entered is negligible. Treatment predicts about 3 to 4% of the variance in posttest raw scores. Analysis of raw scores increases the predictable variance from 55% to 65% and decreases the apparent predictive importance of the treatment factor by about half.

Table 13 shows stepwise regression analyses for Verbal and Reasoning partscores with all grades combined. Predictor variables were IQ partscores on preceding tests, treatment, sex, and grade. (The two grade variables were dummy variables, one contrasting grades

TABLE 13

Results of Stepwise Regression Analyses Using Separate Subscores

Criterion Variable: Separate Subscore IQ Posttests

Predictors: G1 (Grades 1-2 vs. 3-4), G2 (Grades 3-4 vs. 5-6), Sex, Treatment, Preceding IQ Scores

Criterion	Step	Variable entered	F to enter	R^2	Increase in R^2
Verbal IQ 2	1 forced	VIQ 1	409.1	.53	
	2 forced	Treatment	.19	.53	.00
	3 free	Sex and Grade		.54	.01
Verbal IQ 3	1 forced	VIQ 1	427.9	.57	
	2 forced	Treatment	.6	.57	.00
	3 free	VIQ 2	132.2	.70	.13
	4-6 free	Grade and sex		.70	.00
Verbal IQ 4	1 forced	VIQ 1	197.8	.48	
	2 forced	Treatment	4.1	.49	.01
	3 free	VIQ 3	72.4	.62	.13
	4 free	G2	34.2	.67	.05
	5 free	VIQ 2	10.3	.68	.01
	6-7 free	Sex and Grade		.69	.00

TABLE 13 (Continued)

Reasoning IQ 2	1 forced	RIQ 1	159.5	.30	
	2 forced	Treatment	5.7	.31	.01
	3-5 free	Grade and sex		.35	.03
Reasoning IQ 3	1 forced	RIQ 1	106.5	.26	
	2 forced	Treatment	8.4	.28	.02
	3 free	RIQ 2	92.3	.44	.17
	4-6 free	Grade and sex		.46	.02
Reasoning IQ 4	1 forced	RIQ 1	44.6	.18	
	2 forced	Treatment	.95	.18	.00
	3 free	RIQ 3	89.4	.43	.25
	4 free	RIQ 2	29.6	.51	.07
	5-7 free	Grade and sex		.51	.01

1 and 2 with 3 and 4 and the other contrasting grades 3 and 4 with 5 and 6). Pretest IQ was forced into the equation first, and treatment second; the other variables were free to enter in any order. Our previous cautions about interpreting these analyses must be even more strongly emphasized here due to the high frequency of extreme scores in these IQ subscores. For all grades combined, treatment predicts a maximum of 2% of the variance in any IQ subscore. Inclusion of preceding subscores in addition to pretest increased the predictable variance by from 13 to 32%. For Verbal IQ 54%, 70%, and 69% of the second, third, and fourth tests were predictable using all variables; for Reasoning IQ these figures were 35%, 46%, and 51% respectively, providing additional demonstration of the instability of the Reasoning subscores.

Investigation of Treatment Effects Using Analysis of Variance

RJ did not report fully on the analyses of variance performed and did not include any analysis of variance tables. Their only report on actual procedure used is contained in a footnote suggesting they were

> ... following the plan of a multifactorial analysis of variance with interest focused on the main effect of treatments, the two-way interactions of treatments by grades, treatments by tracks, treatments by sex, and treatments by minority-group status. Three-way interactions were also computed for treatments by sex by tracks, treatments by sex by grade levels, and treatments by minority-group status by sex. All other possible three-way and higher-order interactions yielded one or more empty cells or a number of cells with Ns so small as to weaken any confidence in the results even though the analyses were possible in principle.

> All two-way and three-way analyses had unequal and nonproportional Ns per cell, and Walker and Lev's (1953) approximate solution was employed. ...the main effect of treatments was of course obtained in each of the analyses of variance, and p values associated with the F's ranged from .05 to .002. (p. 94-95)†

The Walker and Lev approximate solution referred to by RJ is generally known as "unweighted means analysis."

In this section, we discuss RJ's choice of computation method and their choice of factors to include in the analyses. Later in this section we report the results of several overall analyses of variance as well as some analyses of variance within grade group. These serve primarily to demonstrate how widely the results of slightly different analytic procedures can vary when cell sizes are unequal and data have measurement and sampling problems.

Analysis of Variance in Unbalanced Designs. Application of analysis of variance to problems with unequal cell sizes although common has received too little attention in the literature beyond the cookbook details of computation. When cell sizes are unequal we are faced with several issues: The first and most important question concerns whether analysis of variance still is a valid procedure. Then, if so, what factors should be included? What computational method should be employed?

Standard analysis of variance procedures are based on the assumption that individuals have been assigned at random in equal numbers to each cell of the design (for factors like treatment) or selected at random from a larger group to fill each cell of a cross-classification with an equal number of individuals (for factors like sex). When all cell sizes are equal, the analysis of variance is said to be balanced or orthogonal and the estimates of the various main effects and interactions are orthogonal or statistically independent. If cell sizes in an AxBxC design are all equal, the sums of squares for main effects and interactions of factors A and B are unaffected by the inclusion or exclusion of factor C in the analysis. The only difference between an analysis of variance including only factors A and B and one including factor C also is the size of the error term; generally speaking, the more factors included in the analysis the smaller the error term. Under these circumstances, the full least squares solution with equal weights and the "unweighted means" procedure will produce identical analyses.

If cell sizes in a complete cross-classification were originally equal (or proportional) and subsequent subject losses were equally likely in each cell and thus final cell sizes are not related to the defining factors, an analysis of variance may be performed using the least squares procedure with an appropriate choice of weights. Unweighted means analysis is "a quick approximate analysis to replace the tedious exact calculations" of least squares with equal weights (Scheffé, 1959, p. 362). The adequacy of approximation depends on the amount of variation in cell sizes. With computers so readily available, there seems no justification for using unweighted means analysis. Consequently, we have used the least squares procedure exclusively in our reanalysis.

A major issue is the validity of the analysis of variance approach when cell sizes are related to the defining factors or when collapsing over factors is necessary because cell sizes are zero or very small. Nonrandom cell fluctuations may occur when natural classifications such as intact classrooms are used or when differential subject loss occurs due to treatments. In these situations application of standard analysis of variance procedures may yield misleading results. We illustrate with two examples--one using natural classifications and one involving collapsing of categories. Both illustrate problems which occur in the RJ study.

A simple example based on the interaction in cell size between sex and track observed by RJ illustrates the misleading results an analysis of variance may yield when cell sizes are not independent of factors. Suppose boys and girls were distributed in the three ability tracks as shown in Table 14. Consider two different idealized situations which might produce this situation. In situation A, children are assigned to track strictly on the basis of ability; all children with IQs of 120 are placed in the fast track, all IQs of 100 are placed in the medium track, all IQs of 80 are placed in the slow track. Thus, to produce the cell sizes shown,the IQ distribution by sex must be that shown under situation A; the resulting cell means are also shown. In situation B, boys and girls have the same IQ distribution but girls are more likely to be placed in fast or medium tracks than boys. Thus not only are all the girls with IQs of 120 placed in the fast track, but also 20 of the girls with IQs of 100 are placed in the fast track, giving a cell mean for girls in the fast track of (30x120 + 20x100)/50 = 112. Conversely only 20 of the 30 boys with IQs of 120 are placed in the fast track, the rest are placed in the medium track and so on.

Applying the least squares procedure with equal weights we obtain a main effect for track in both situations. However, in situation A we would obtain no sex effect and no sex x track interaction. In situation B, we would obtain a sex effect and a track x sex interaction. Thus, in both situations an analysis of variance produces misleading conclusions about IQ differences between the sexes.

Next we illustrate the misleading results that can be obtained when factors are dropped from an unbalanced design. In Table 15 is an idealized example of cell sizes for treatment x track in one grade--these figures are very similar to those actually obtained by RJ (see Table 2). Suppose that there is really no treatment effect but that children in the fast and slow tracks tend to gain more than children in the middle track and that we obtain the mean gains shown. When least squares with equal weights is applied to the treatment x track classification we obtain no treatment main effect and no treatment x track interaction. Suppose, however, that it was decided to omit the track factor because of small sample size or to allow introduction of sex as a factor, then, due to the unbalanced sample sizes, we would obtain a spurious treatment effect.

Although RJ assigned children to the experimental and control groups to produce cell sizes in the ratio of about 1 to 4, they used an unweighted analysis; every cell was assigned equal weights in the calculation of main effects and interactions. If there are no interactions, the results are unaffected by the choice of weights and the standard procedure is to choose equal weights. If there is interaction, tests for main effects will be affected by the choice of weights. If the control group receives a weight of 4 and the treatment group a weight of 1 and all other effects are defined using equal weights,then the main effect for treatment and all interactions involving treatment will be the same as if equal weights were used; all other main effects and interactions will be affected by the choice of weights. Since there is no compelling

TABLE 14

Example of Two Idealized Situations Producing an Interaction in Sex x Track Cell Sizes

Cell Size

Track

		Fast	Medium	Slow	
Sex	M	20	30	50	100
	F	50	30	20	100
		70	60	70	200

Situation A

IQ Distribution
Number of Children
IQ

		120	100	80
Sex	M	20	30	50
	F	50	30	20

Cell Means

		Fast	Medium	Slow
Sex	M	120	100	80
	F	120	100	80

Situation B

IQ Distribution
Number of Children
IQ

		120	100	80
Sex	M	30	40	30
	F	30	40	30

Cell Means

		Fast	Medium	Slow
Sex	M	120	106.7	88
	F	112	93.3	80

Actual Cell Sizes for Third and Fourth Graders at Basic Posttest

Track

		Fast	Medium	Slow
Sex	M	13	19	24
	F	30	16	13

TABLE 15

Idealized Example Showing the Effect of Dropping Factors

Cell Sizes

		Track		
		Fast	Medium	Slow
Treatment	C	15	15	15
	E	5	1	5

Mean Gains

		Track		
		Fast	Medium	Slow
Treatment	C	1.0	0.0	1.0
	E	1.0	0.0	1.0

reason to calculate sex and grade effects as if the experimental and control groups were equal in size, we decided to calculate most of the analyses of variance using a least squares analysis with proportional weights. The F tests for treatment and interactions with treatment will be the same with proportional weights as with equal weights but the calculated effects for sex, grade, and track will be much more heavily influenced by the larger control group using proportional weights.

The following technical discussion illustrates this point. Consider a two-way layout with possible interaction. The model is

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}$$

where

$$i = 1, 2, \ldots, I, \quad j = 1, 2, \ldots, J$$

so we have an I x J layout. This model for the cell means contains $1 + I + J + IJ$ parameters, but there are only IJ cells and therefore only IJ parameters can be estimated. So we must impose conditions on the parameters. These conditions can be identified as follows:

1) Select a set of weights corresponding to the levels of A, $\{u_i\}$ where $u_i \geq 0$ and $\Sigma u_i = 1$, and a set of weights corresponding to the levels of B, $\{w_i\}$ where $w_i \geq 0$ and $\Sigma w_i = 1$.

2) Then impose conditions

$$\sum_i u_i \alpha_i = 0$$

$$\sum_j w_j \beta_j = 0$$

$$\sum_i u_i \gamma_{ij} = 0 \text{ all } j \qquad \sum_j w_j \gamma_{ij} = 0 \text{ all } i$$

With these conditions, the mean of the i^{th} level of A is $A_i = \sum_j w_j \mu_{ij}$, the mean of the j^{th} level of B is $B_j = \sum_i u_i \mu_{ij}$, and we define $\mu = \Sigma\Sigma\, u_i w_j \mu_{ij}$, and $\gamma_{ij} = \mu_{ij} - B_j - A_i + \mu$.

If in fact $\gamma_{ij} = 0$ for all i, j (no interaction), then the choice of weights will not affect SS_A or SS_B or any contrast among the α_i or β_j. Therefore, if there is no interaction, it will not matter what weights are chosen; the standard procedure would be to choose equal weights. If there is an interaction, the test of SS_{AB} is unaffected by the choice of weights but the main effects and tests on SS_A and SS_B will depend on the weights chosen.

An example will show what happens to the sums of squares for A and the sums of squares for B when we use unweighted means analysis, least squares with equal weights, and least squares with proportional weights (choosing $u_1 = u_2 = 1/2$, $w_1 = 5/6$, $w_2 = 1/6$). For a particular case where

Cell Sizes

n_{ij}

	B_1	B_2
A_1	10	2
A_2	10	2

Cell Means

$\bar{x}_{ij}$

	B_1	B_2
A_1	10	22
A_2	10	10

Unweighted means	$SS_A = 120$	$SS_B = 120$
Least squares with equal weights	$SS_A = 120$	$SS_B = 120$
Least squares with proportional weights	$SS_A = 24$	$SS_B = 120$

Thus, in estimating the effect of A, the cell with a mean of 22 receives much less weight when we take account of its small sample size by using proportional weights. The conclusion about B is unaffected by the use of proportional weights. Unweighted means and unweighted least squares give the same results; they would not if cell sizes were not exactly but only approximately proportional.

Results of Analyses of Variance. We computed several overall analyses of variance using Total IQ pretest and Total IQ posttests as criterion variables. Two analyses of Total IQ gain scores were included for comparison with RJ's computations. Results are shown in Table 16. For completeness, the same analyses were computed for verbal and reasoning subscores, although interpretation of these results is doubtful (see Tables 17 and 18). Separate analyses of variance were computed within each grade group with posttest as criterion, gain scores as criterion, and posttest with pretest as a covariate (see Tables 20-22). These analyses of variance allow us to compare the results obtained with different choices of factors, different criterion measures, different sets of weights and different treatment of extreme scores.

Our discussion of analysis of variance in unbalanced designs illustrates how important the choice of factors is to the results obtained. Ideally treatment, track, grade, sex and minority group should all be included as factors in the analysis. This is impossible. Consequently some factors must be dropped or factors such as grade must be reduced from 6 levels to 3. Decisions about how to reduce the number of factors must be guided by the sampling and balancing needs of the design as well as by the purposes of the experiment.

We have dropped the minority group factor from our analyses of variance. The Mexican vs. non-Mexican factor was not a part of the design of the experiment; other variables describing ethnic origin or socio-economic background could as easily have been analyzed. Since only 17% of the children were Mexican and this factor interacts with sex and track in cell size, its introduction sharply reduces cell sizes and it is unclear that a satisfactory assessment of its significance could be made.

Retaining grade, track, and sex there are still too few children per cell; there are 72 cells of which 6 are empty and many have only 1 or 2 children. As noted earlier, there are more girls in the high track and more boys in the low track so analyses of variance including both sex and track would likely produce misleading conclusions about the effects of these variables.

The children in grades 1 and 2 both received TOGA Form K-2, those in grades 3 and 4 received Form 2-4, and those in grades 5 and 6 received Form 4-6. Since RJ combined these grades for some analyses, it seemed reasonable to use grade group rather than grade in some of our analyses to improve cell size.

Tables 16 through 18 summarize the results of analyses of variance with three choices of factors: treatment by grade group by sex (TxG'xS), treatment by grade by ability track (TxGxA), and treatment by grade group by sex by ability track (TxG'xSxA). Treatment by grade by ability track is the same as treatment by classroom and is probably the most important single analysis. For the basic posttest grade 5 had to be deleted because classroom 5B did not take the Reasoning subtest. The other two analyses both contain treatment by grade group by sex and comparison of their results shows what happens when the factor of ability track is included or excluded.

Analyses were performed on IQ scores from all four testings and on gain from pretest to basic posttest. Some analyses used all data, others truncated data; all were done using least squares, some using equal weights and some using proportional weights. Note that none of these analyses reproduce exactly any of those performed by RJ. Effects significant at the .05 level are indicated in the tables; blank cells in the tables indicate analyses not performed.

Total IQ is the only measure sufficiently reliable to admit interpretation. Looking at the results for pretest Total IQ we gain a consistent picture of grade and ability track differences. Note, also, the triple interaction involving treatment. Results for Total IQ at second testing show how the presence of a sex effect is affected by the treatment of extreme scores.

Analyses of Total IQ basic posttest fairly consistently indicate some treatment effect although with the consistent superiority of the experimental group on the pretest these results can only be regarded as suggestive that further more carefully chosen analyses should be undertaken. The fact that inclusion of more factors or exclusion of extreme scores reduces the treatment main effect to a three-way interaction is an indication that treatment effects are probably present in only a few cells of the classification.

The two analyses performed using gain scores with all the data and equal weights should provide results closest to those obtained by RJ. It is interesting to note that the only consistent results obtained in these two analyses is a grade effect. RJ may have obtained significant treatment effects in every analysis but we do not.

The consistent appearance of grade main effects and interactions involving grade confirms our earlier contention that separate analyses be made for different forms of TOGA (or grade groups).

Although we do not recommend analysis of verbal and reasoning partscores, we note that these analyses provide no indication whatever

TABLE 16

Analysis-of-Variance Results: Verbal IQ

Effects Significant at .05 Listed

Criterion	Weights	Data Set		Factors	
			TxG'xS	TxGxA††	TxG'xSxA
Total IQ 1	E	All		G,A,GxA TxGxA	G,A
	P	All	G'		
	P	Truncated	G'		
Total IQ 2	P	All	T,G',S		
	P	Truncated	T,G'		
Total IQ 3	E	All	T	T,A,GxA	A,G'xSxA, TxG'xS
	P	All	T,G'		
	E	Truncated	TxG'xS		
	P	Truncated	G',TxG'xS		
Gain TIQ3-TIQ1	E	All		T,G	G',G'xA
Total IQ 4	P	All	S		
	P	Truncated	T,G'		

P = proportional weights
E = equal weights
TT = both pretest and posttest of interest truncated
A = denotes track or ability grouping
G' = the three grade levels--one and two, three and four, and five and six
††= Grade 5 has been deleted from this analysis because classroom 5-B did not take the Reasoning subtest

TABLE 17

Analysis-of-Variance Results: Total IQ

Effects Significant at .05 Listed

Criterion	Weights	Data Set		Factors	
			TxG'xS	TxGxA††	TxG'xSxA
Verbal IQ 1	E	All		A,GxA	A
	P	All	G',S		
	P	Truncated	*		
Verbal IQ 2	P	All	S		
	P	Truncated	S		
	P	TT	S		
Verbal IQ 3	E	All		A	A
	P	All	S		
	P	Truncated	*		
	P	TT	*		
Gain Verbal IQ 3- Verbal IQ 1	E	All		GxA	G'xA
Verbal IQ 4	P	All	G',S		
	P	Truncated	G',S		
	P	TT	G',S		

P = proportional weights
E = equal weights
TT = both pretest and posttest of interest truncated
A = denotes track or ability grouping
G' = the three grade levels--one and two, three and four, and five and six
*There were no effects significant at .05
††= Grade 5 has been deleted from this analysis because classroom 5-B did not take the Reasoning subtest

TABLE 18

Analysis-of-Variance Results: Reasoning IQ

Effects Significant at .05 Listed

Criterion	Weights	Data Set		Factors	
			TxG'xS	TxGxA††	TxG'xSxA
Reasoning IQ 1	E	All		G,A	G',A
	P	All	G',S		
	P	Truncated	G',S, G'xS		
Reasoning IQ 2	P	All	T,G'		
	P	Truncated	T,G'		
	P	TT	T,G'xS		
Reasoning IQ 3	E	All		T,A	G,A,TxG'xS TxG'xA G'xSxA
	P	All	T,G', G'xS		
	P	Truncated	G', G'xS		
	P	TT	G', G'xS		
Gain	E	All		G	G',TxS
Reasoning IQ 4	P	All	G'xS		
	P	Truncated	G'xS		
	P	TT	G'xS		

P = proportional weights
E = equal weights
TT = both pretest and posttest of interest truncated
A = denotes track or ability grouping
G' = the three grade levels--one and two, three and four, and five and six
†† = Grade 5 has been deleted from this analysis because classroom 5-B did not take the Reasoning subtest

TABLE 19

Analysis for Decrease in Error Variance Due to Use of Gain Scores

		Factors	
		TxGxA	TxG'xSxA
Total IQ	Error Variance using Posttest	243	243
	Error Variance using Gain	155	166
	Decrease in Variance	36%	32%
Verbal IQ	Error Variance using Posttest	649	629
	Error Variance using Gain	316	321
	Decrease in Variance	51%	49%
Reasoning IQ	Error Variance using Posttest	584	627
	Error Variance using Gain	610	714
	Decrease in Variance	-4%	-14%

of a treatment effect on the verbal subtest. Our analyses of reasoning gain do not confirm RJ's report of very significant main effects and the treatment effects which do appear for Reasoning IQ basic posttest disappear when extreme scores are removed.

Table 19 provides a summary of the relative precision of gain scores versus posttest scores obtained from analyses reported in Table 16. These analyses were calculated using least squares with equal weights on all the data.

Turning now to separate analyses by grade group, Tables 20-22 provide comparisons of results obtained using pretest gain scores, posttest only, and posttest with pretest as a covariate. Sex and track were not included in the analyses. Results are shown in terms of "expectancy advantage," that is, mean difference between experimental group and control group scores. Calculations were repeated on renormed and truncated IQ scores as well as raw scores for 1st and 2nd graders. (Pretest and posttest were jointly renormed or truncated.)

Examining Table 20 for Total IQ, we note that the three criterion measures and three sets of scores consistently show no expectancy advantage for third, fourth, fifth, and sixth graders. Results for first and second grades do seem to indicate an expectancy advantage but we note the 4 to 5 point advantage on the pretest and our earlier uncertainty that any of these analyses could be regarded as valid. These results warrant a closer look at first and second graders and further attempts to construct a valid analytic procedure in the face of pretest advantage, unreliability, and imbalance. Notice that renorming and truncation tend consistently to reduce apparent differences between the experimental and control groups.

Analyses of Verbal IQ and Reasoning IQ partscores are generally consistent with the results obtained for Total IQ. Note, however, how widely the apparent results differ depending on the treatment of extreme scores and the selection of criterion.

Analysis by Classroom

In our analyses to this point, we have treated the individual child as the experimental unit. What happens if the classroom is considered

TABLE 20

Pretest to Basic Posttest "Advantage" in Total IQ

Mean scores for experimental group minus mean scores for control group

	Pretest	Posttest	Gain	Posttest adjusted for pretest
Grade Group				
First and Second Grades				
All IQ	4.9	15.9*	11.0*	12.8*
Renormed IQ	4.5	13.7*	9.2*	10.8*
Truncated IQ	0.7	10.6*	9.9*	10.1*
Raw Scores	4.0	6.5*	2.5	4.4*
Third and Fourth Grades				
All IQ	0.5	2.3	1.8	2.0
Renormed IQ	0.5	2.1	1.6	1.6
Truncated IQ	-1.9	0.1	2.0	1.7
Fifth and Sixth Grades				
All IQ	4.3	4.5	0.2	- 0.1
Renormed IQ	4.3	4.4	0.1	0.1
Truncated IQ	3.6	2.3	-1.3	-1.4

*Two tailed $p < .05$

TABLE 21

Pretest to Basic Posttest "Advantage" in Verbal IQ

Mean scores for experimental group minus mean scores for control group

Grade Group	Pretest	Posttest	Gain	Posttest adjusted for pretest
First and Second Grades				
All IQ	0.4	10.5*	10.1*	10.2*
Renormed IQ	0.5	9.0*	8.5*	8.7*
Truncated IQ	-1.4	6.9	8.3*	7.8*
Third and Fourth Grades				
All IQ	4.0	-0.6	-4.6	-4.8
Renormed IQ	3.2	-3.6	-6.8*	-6.4*
Truncated IQ	-1.7	-7.3	-5.6	-5.9
Fifth and Sixth Grades				
All IQ	0.7	2.7	2.0	2.0
Renormed IQ	0.7	1.0	0.3	0.4
Truncated IQ	3.0	1.6	-1.4	-1.0

*Two tailed p < .05

TABLE 22

Pretest to Basic Posttest "Advantage" in Reasoning IQ

Mean scores for experimental group minus mean scores for control group

	Pretest	Posttest	Gain	Posttest adjusted for pretest
Grade Group				
<u>First and Second Grades</u>				
All IQ	13.2	25.8*	12.6	21.0*
Renormed IQ	8.4	18.6*	10.2	13.7*
Truncated IQ	0.3	6.0	5.7	5.8
<u>Third and Fourth Grades</u>				
All IQ	-3.0	5.7	8.7	8.3
Renormed IQ	-3.0	6.3	9.3*	8.5*
Truncated IQ	-3.4	6.9	10.3*	9.0*
<u>Fifth and Sixth Grades</u>				
All IQ	4.0	8.9	4.8	5.4
Renormed IQ	4.1	3.9	-0.2	0.5
Truncated IQ	3.2	-1.6	-4.8	-4.0

*Two-tailed p < .05

to be the unit of observation? Expectancy effects are after all probably group phenomena. The test information is provided to a teacher who in turn operates on a whole classroom. Although eventually to be detected in individual student performance, expectancy effects may best be understood as a function of the particular groups in which they occur. There is, then, much justification for considering the experiment as a sample of 18 classrooms each with a subgroup of experimental and control subjects.

RJ applied the _t_ test, the Wilcoxon and the sign test to the eighteen pairs of mean gains. We also want to investigate pre and posttest means. The sample size of experimental and control groups varies widely from classroom to classroom and there are fairly sizeable IQ differences between grades and between tracks. As a consequence, RJ's application of the _t_ test and the Wilcoxon test is inappropriate, since both require that difference scores for each pair represent a random sample from _one_ distribution. If we can assume that assignment to treatment was random and that no differential selection bias occurred, the sign test can be employed to test the null hypothesis that in any classroom the probability of the experimental group having a higher mean (or higher gain) than the control group is one half $(P\ (E > C) = 1/2)$.

Pretest means used here were for those individuals present at the basic posttest. Classroom means for basic posttest and gains are taken from RJ Tables A-4 to A-9. They thus include all extreme scores. For thoroughness the sign test analyses we report should also be performed for means of the truncated data. Classroom 5B had no posttest reasoning scores and was deleted where necessary.

TABLE 23

Analysis by Classroom: Total IQ

Total IQ Scores

			Pretest	Posttest	Gain
#Classes	E > C		9	13	11
	E < C		8	4	6
		Total	17	17	17
Two tail p			1.0	.04	.34

Change from pre to posttest

		# Classes Posttest		
		E > C	E < C	
Pretest				
	E > C	8	1	9
	E < C	5	3	8
		13	4	17

TABLE 24

Analysis by Classroom: Verbal IQ and Reasoning IQ

Verbal IQ

		Pretest	Posttest	Gain
#Classes	E > C	11	11	12
Total		18	18	18
Two tail p		.48	.48	.24

Reasoning IQ

		Pretest	Posttest	Gain
#Classes	E > C	12	13	15
Total		17	17	17
Two tail p		.14	.04	.002

Change from pre to posttest

		Post		
		E > C	E < C	
Verbal IQ	Pretest			
	E > C	8	3	11
	E < C	3	4	7
		11	7	18
Reasoning IQ				
	E > C	11	1	12
	E < C	2	3	5
		13	4	17

For Total IQ (see Table 23) there are a total of 17 classrooms; the experimental group gained more than the control group in eleven--not significantly more than half of the classrooms. The experimental group did have a higher posttest mean in 13 classrooms but looking at changes in ranking from pre to posttest we note that in eight of these classrooms the experimental group was higher to begin with. Verbal IQ shows no significant evidence of experimental group superiority. For Reasoning IQ, eleven of the classrooms were superior on both pre and posttest.

A Closer Look at First and Second Graders

We have examined the results of many different analyses. For the third through sixth grade we conclude that there is no evidence of a treatment effect. Results for first and second graders, however, are inconclusive. Although the application of standard statistical procedures yields significant differences in treatments, the doubtful measurements and uncertain sampling procedure and balance make it unclear whether any of the analyses are valid. As a consequence we must take a closer look at total raw scores for these children. Using raw scores does not take differences in age into account but the stepwise regression reported in Table 12 indicates that age is essentially unrelated to raw score gain for this group anyway. Table 25 shows the ages and pretest and posttest raw scores for first and second grade children grouped by sex and classroom. Control group children are listed according to rank on the pretest; each experimental group child is shown beside that control group child whose pretest score provides the closest match. (There are 95 control children and 19 experimental children.)

TABLE 25

Pre and Posttest Raw Scores for First and Second Graders

First Grade Track 1

Male						Female					
Age	Pre	Post				Age	Pre	Post	Age	Pre	Post
Control			Experimental			Control			Experimental		
6.3	11	40	5.5	10	41.5*	6.2	7	26			
6.0	13	39				5.6	10	28	5.7	10	37
6.3	14	28				5.9	11	30			
6.0	15	37				6.3	18	34			
5.6	16	53				6.0	20	31.5			
5.8	20	41.5				6.4	21	33			
5.8	21	26									
6.2	23.5	41.5									
5.6	27	35									
6.0	39	45									

First Grade Track 2

Male						Female					
Control			Experimental			Control			Experimental		
6.0	5	44				5.7	22	31.5	5.7	19	44
5.8	9	37				6.2	22	27	5.7	20	52
5.8	15	31.5				6.3	23.5	41.5			
6.3	20	35				5.9	29	38			
5.6	21	41.5				5.8	35	43			
5.7	22	36				5.5	37	49			
5.7	23.5	45									
5.7	23.5	46									
5.6	27	49	5.5	26	43						
			6.0	41.5	56						

*Sometimes two different raw scores corresponded to the same mental age; in converting IQ scores back to raw scores in these cases, the average of the two raw scores was used.

TABLE 25 (Continued)

First Grade

Track 3

Male						Female					
Age	Pre	Post				Age	Pre	Post	Age	Pre	Post
Control			Experimental			Control			Experimental		
6.1	22	44				6.0	27	49			
5.5	23.5	44				6.2	27	39			
5.9	25	43				6.3	28	43			
6.0	31.5	53				5.7	29	52			
5.7	39	53				6.1	34	31.5	6.4	33	51
5.7	41.5	45				5.5	36	50			
6.4	41.5	55				6.2	38	38			
6.3	43	55									
6.2	44	48									
6.5	51	54									

Second Grade

Track 1

Male						Female					
Control			Experimental			Control			Experimental		
7.2	17	41.5				7.2	22	26			
7.9	17	44				6.7	23.5	30			
			6.6	25	47	6.8	23.5	38			
6.7	31.5	37	7.5	31.5	48	7.2	25	45			
			8.1	31.5	48	6.9	26	39			
7.9	41.5	52				6.9	30	46			
						7.1	31.5	48			
						7.5	31.5	51			
						6.9	33	41.5			
						6.9	36	49			

TABLE 25 (Continued)

Second Grade

Track 2

Male						Female					
Age	Pre	Post				Age	Pre	Post	Age	Pre	Post
Control			Experimental			Control			Experimental		
6.9	23.5	40				8.0	30	35			
6.7	26	46				7.1	31.5	51	6.8	33	43
7.2	26	41.5				7.1	49	57	7.0	46	59
7.0	31.5	53									
6.5	33	49									
7.3	33	51									
7.4	33	51									
6.9	35	49									
6.6	36	43									
7.3	41.5	44									
6.8	46	57	6.5	44	49						

Second Grade

Track 3

Male						Female					
Control			Experimental			Control			Experimental		
6.7	36	50				6.7	29	53			
6.7	40	55				7.3	40	40			
7.4	43	56				7.0	41.5	49			
7.5	46	55	6.9	46	57	7.0	41.5	51			
6.8	48	55				6.7	41.5	44	6.6	41.5	56
7.2	49	58				6.9	41.5	50			
6.5	50	57				7.2	43	50	7.4	45	56
7.1	50	58				6.8	47	47	6.8	45	59
7.4	50	56				7.4	53	55			
6.5	51	56	7.2	53	56						
			7.2	56	63						

The attempt to find a comparable control group child of the same classroom and sex to match with each experimental group child reveals several things. First, there were four experimental children who could not be matched because there was no control group child with a pretest score within ± 3 points. Second, in the twelve cells there were two with no experimental child at all, 4 with one, 3 with two, and 3 with three children. Eleven of the experimental children were young in comparison with the control group, seven of these were the youngest in their group; four were old in comparison with their group, two being the oldest. Thus 16 of the 19 experimental group children were extreme in age in comparison with classmates of the same sex, and 9 were the most extreme.

Looking at pretest scores in the same way we find four experimental group children with low pretest scores, three with the lowest; seven experimental children with high pretest scores, three with the highest. Thus six of the experimental group children had pretest scores which were either the highest or the lowest among classmates of the same sex. We thus obtain somewhat clearer evidence that the control and experimental children do not provide closely comparable groups. It is therefore unclear whether any analysis can clarify the issue of whether or not there is a treatment effect. We may, however, gain some insight by looking further at the scores of the two groups.

First we examine raw score gains for the matched children (see Table 26). We note that reasonable matches were obtained only for 15 of the 19 experimental children. Looking at signs only we find 3+, 3- for boys and 8+, 1- for girls for a total of 11+, 4-. Using the sign test then there is no significant difference in gains between the pairs ($p > .05$ one sided). Using a Wilcoxon signed rank test, we obtain sum of negative ranks = 24 which is significant at .05. The median "excess gain" was 5. Since the magnitude of gain in raw score which is possible depends on the pretest score and thus varies considerably from grade 1 slow track to grade 2 high track, the t-test on gains does not seem a valid choice and the Wilcoxon signed rank test is also of dubious validity.

Looking at gain in relative rank for each experimental child in comparison with his classroom and sex group (e.g. for males in grade 1 track 1 the experimental child ranks lowest on the pretest but ranks eighth on the posttest for a change in rank of +7) we obtain two zero changes, four negative changes, and 13 positive changes. These results would be significant at the .05 level using the sign test. This analysis does not allow for the fact that individuals below the median on the pretest can be expected to have positive rank changes. Table 27 shows that 6 experimental children showed changes in rank from below to above the median and 1 showed a downward change; this is not significant.

Suppose we look at the problem a different way. If the treatment were effective we ought to be able to distinguish between experimental and control group children on the basis of posttest or gain scores. Can we do so? How successfully can children be classified as being from the experimental or control group on the basis of posttest or gain scores alone? For example, there is one experimental boy in grade 1, track 1; if we pick the boy with the highest posttest score from the eleven boys in grade 1, track 1, will it be the experimental child? Results using highest posttest scores are shown in Table 28 and using highest gain in Table 29.

TABLE 26

Excess of Gain by Experimental Children for the 15 "Matched" Pairs

	Sex	
	Male	Female
Grade 1		
Track 1	2.5	9
2	-5.0, --	15.5, 27
3	---	20.5
Grade 2		
Track 1	11, --, --	--
2	-6	-9.5, 5
3	2, -2, --	7.5*, 4, 14

*This experimental girl could have been matched with any of four control group children yielding "excess gains" of 12, 7, 6, 5; we have computed the average.

TABLE 27

Changes in rank within sex and classroom

	Posttest		
	Below median	Above median	
Pretest			
Below median	2	6	8
Above median	1	10	11
	3	16	19

TABLE 28

Children with Highest Post Score

Male

	No. of Control Children	No. of Experimental Children	Identity of Those Selected	No. Actually Experimental
Grade 1				
Track 1	10	1	C	0
2	9	2	C, E	1
3	10	0	---	---
Grade 2				
Track 1	4	3	C, E, E	2
2	11	1	C	0
3	10	3	C, C, E	1

E = 2.5

Female

	No. of Control Children	No. of Experimental Children	Identity of Those Selected	No. Actually Experimental
Grade 1				
Track 1	6	1	E	1
2	6	2	C, E	1
3	7	1	C	0
Grade 2				
Track 1	10	0	---	---
2	3	2	C, E	1
3	9	3	E, E, E	3

		Children Actually		
		E	C	
Classified as	E	10	9	19
	C	9	86	95
		19	95	114

TABLE 39

Children with Highest Gain Scores

Male

		No. of Experimental Children	Identity of Those Selected	No. Actually Experimental	
Grade 1					
Track	1	1	C	0	
	2	2	C, C	0	
	3	0	---	---	
Grade 2					
Track	1	3	E, C, C	1	
	2	1	C	0	
	3	3	C, C, C	0	E = 2.5

Female

		No. of Experimental Children	Identity of Those Selected	No. Actually Experimental
Grade 1				
Track	1	1	E	1
	2	2	E, E	2
	3	1	C	0
Grade 2				
Track	1	0	---	---
	2	2	E,C	1
	3	3	E, E, C	2

Children Actually

		E	C	
Classified as	E	7	12	19
	C	12	83	95
		19	95	114

Using highest posttest score, we correctly classify 10 of the 19 experimental children; using pretest we would identify 7; 5 are highest on both pre and posttest. Using highest gain score we correctly classify 7 of the 19 experimental children. In either case, the expected number of experimental children correctly classified by selecting at random is 4.8 with a standard deviation of 1.65. Using gain scores then we do not correctly classify more experimental children than we would expect to by selecting at random. (See Appendix A, p. 145.)

Our closer look at first and second graders using raw scores to test for differences between experimental and control children has produced mixed results. The small sample size and lack of balance make it difficult to find a really appropriate analytic procedure. There are indications that the control and experimental group children are insufficiently comparable to make any sound conclusions. Examination of the data suggests that there is no expectancy effect for boys but that there may be one for girls.

In conclusion then there is some evidence to suggest the presence of an expectancy effect in first and second graders. However, with so small and poorly balanced a sample, a conclusive analysis of these data is not possible. Definitive conclusions require additional experiments.

CHAPTER VI: CONCLUSIONS

The Pygmalion Effect

Our reanalysis reveals no treatment effect or "expectancy advantage" in Grades 3 through 6. The first and second graders may or may not exhibit some expectancy effect; these experimental and control groups differ greatly on the pretest and a statistical analysis of such data cannot provide clear conclusions. There is enough suggestion of an expectancy effect in Grades 1 and 2 to warrant further research, but the RJ experiment certainly does not demonstrate the existence of an expectancy effect or indicate what its size may be.

Experimenters continuing work in this area should make strenuous efforts to obtain more precise measurement and more carefully controlled experimental treatments. More recent investigations have attempted to study expectancy effects in teachers. Since most of this work is as yet unpublished, it is difficult to know whether significant improvements in technique have been made. Rosenthal (1969a, 1969b) has summarized a number of these studies and concluded that they provide strong combined evidence of teacher expectancy operating to influence student learning. Meanwhile, Rosenthal's (1966) earlier lines of laboratory research on experimenter bias have been severely criticized by Barber and Silver (1968) and both our review and that by Claiborn (1969) show that many of these earlier difficulties have been carried forward into research on teacher expectancy. There are signs, however, that other investigators are modifying the techniques of earlier research. The recent study by Claiborn improved significantly on the original design and analysis plan, while including enough of the key features of the RJ work to serve as a replication. Claiborn's results were negative; neither total nor subtest IQ showed significant expectancy effects. Although Claiborn's study differs from RJ's in some important respects and although it does not overcome some significant problems in the RJ work identified here, it does take a step in the right direction. It remains to be seen whether other studies will confirm or deny what can at present only be regarded as an intriguing hypothesis.

Recommendations for Further Research

As an aid to planning further research on teacher expectancy effects, as well as a summary of the present report, we close with a brief review of recommendations for consideration by future investigators.

1. As a first step in planning research, state as clearly as possible the proposition under study. This statement should suggest immediately what the key features of the research design are to be. Comparison of proposition and plan will show if questions other than the stated one are implied by the design. For example, RJ (p. 61)[†] stated that their experiment "... was designed specifically to test the proposition that within a given classroom those children from whom the teacher expected greater intellectual growth would show such greater growth." However, RJ did not really plan their primary analyses to be conducted "within classrooms" and never asked the teachers to indicate "those children from whom they expected greater intellectual growth."

2. Define as clearly as possible the psychological construct being measured. Avoid questionable connotations in naming variables. Consider in detail the scale of measurement, the reliability, and the

construct validity of the measures chosen, whether they represent independent or dependent variables. Provide at least two separate measures of all constructs of primary interest in the experiment and examine the extent to which the data support or qualify the original formulation of the construct in question. RJ frequently used terms like "intellectual growth" and "expectancy advantage" in referring to their dependent variable, never discussing the possibility that their simple IQ gain score might not represent the construct of interest to them. RJ offered no information about raw scores or mental ages on their single instrument and made no direct use of other intellectual measures, some of which must have been available from school records. "Intellectual growth" must mean more than changing a few answers the second time through a single test. Neither the reliability nor the validity issues involved in this measure were fully explicated or studied. The term "expectancy advantage" also presumes interpretations before effects are found, a practice especially to be condemned in publications like Pygmalion which are aimed directly at the lay public. Words like "special" and "magic" are frequently used by RJ to refer to experimental children, when less imaginative words would serve as well.

3. Specify as clearly as possible the population to which generalization is planned. Spell out in detail the steps involved in the sampling plan. Where alternative procedures for sampling or assigning subjects to experimental conditions exists, or where subjects are excluded from the analysis, summarize the reasoning that led to the decisions made. After producing a preliminary design, list all possible alternative interpretations for alternative expected results. Modify or expand

the design to eliminate competing and confounding hypotheses and clarify in simple terms the outcomes expected and the implications of each outcome for the hypothesis of interest. Avoid unnecessarily complex designs and the addition of variables of marginal relevance. The final sampling plan and design should provide clear balance with respect to the main comparisons planned. RJ actually said little about the sampling plan. The need for balancing and the effect of its loss were not made clear. The reader was left uncertain regarding many points of concern regarding subject loss, transfer and balancing, and the effects of these issues on the results.

4. Validate the experimental treatment by providing checks and observations to ascertain that treatments really represent for the subjects what they were planned to represent for the experimenters. Observe and describe subject behavior in test administration conditions as well as in experimental treatment conditions. RJ could have included observations of teachers and students during tests and teaching but chose not to do so. The teacher interview, on the other hand, was a useful addition. It showed, however, that RJ's teachers could not remember, and perhaps had never known, who the "bloomers" were in the first place.

5. Look carefully at the basic raw data, before applying complex scoring formulae, transformations, or summarizations. Plot all relationships of interest graphically. One picture is worth many summary numbers. Use simple statistical computations to probe the assumptions and adequacy of more complex statistical abstractions. The most appropriate and productive mental set for the experimenter is that of a

detective, not a defense attorney. Analyze the data in several alternative ways. RJ gave no evidence of having looked at raw data, scatterplotted relations, or probed into the structure of their analyses. Alternative methods of analysis were not discussed and the adequacy of the methods chosen was not questioned.

6. Emphasize the strength and character of relationships. Avoid reducing continuous variables to dichotomous conceptualizations and decisions. Consider the amount of criterion variance accounted for in a relation at least as important as its statistical significance. Report p values within any predetermined limits, but interpret no relation unless $p < .05$. Report p values less than .01 as " < .01." RJ relied almost completely on significance tests to characterize the importance of their findings and wrongly used p value as a measure of strength of effect to indicate size and practical significance of mean differences. Nominal p-values ranging from .25 to .00002 were quoted throughout their work.

7. Use the full power of the data to reach simple rather than complex conclusions, whenever the former account for the data. The form of analysis chosen by RJ led them into unnecessarily complex results. Forming gain scores does not use the power of the data; using IQ instead of raw scores adds to complexity. Treating the four test occasions in separate analyses ignores the powerful repeated measures aspect of the data. Analyzing reasoning, verbal and total scores separately also adds to complexity, since the latter is a simple summation and thus is literally dependent upon the first two subscores. RJ conducted many separate analyses without attempting to show the full set of possible

comparisons or to use interrelationships among variables for data reduction. Their unweighted means analysis is a gross approximation to least squares solutions at best, especially when proportional cell sizes were expressly built into the experiment.

8. Report the results of research as fully and as clearly as possible, using appendices and supplementary publication where necessary. Use scientific and professional journals as the initial outlet for research findings, paying conscientious attention to the suggestions and criticisms of referees and reviewers. Single unreplicated studies of broad public concern should not be reported directly to the public. Incorporate findings into popular books only with due regard for the degree of their possible substantiation by other research and their possible misinterpretation by the public.

The educational researcher will deal increasingly with hypotheses and conclusions of far reaching social importance. While researchers are always responsible for the proper conduct and reporting of research, nowhere should this responsibility be more keenly felt and exercised than in work bearing directly on urgent and volatile social issues. It is essential, then, that both researchers and publishers recognize this responsibility and pursue it to the utmost. It is hoped that this report will help to equip future workers for that pursuit.

REFERENCES

Aiken, L. Review of Pygmalion in the Classroom. Educational and Psychological Measurement, 1969.

Barber, T. X. & Silver, M. J. Fact, fiction, and the experimenter bias effect. Psychological Bulletin Monographs, 1968, 70(6, Part 2), 1-29.

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

Cochran, W. G. Errors of measurement in statistics. Technometrics, 1968, 10, 637-666.

Coles, R. What can you expect? The New Yorker, April 9, 1969, 169-177.

Cronbach, L. J. & Furby, L. How should we measure "change"--or should we? Psychological Bulletin, Spring 1970, in press.

Cronbach, L. J. & Snow, R. S. Individual differences in learning ability as a function of instructional variables. Final Report, USOE Contract No. 4-6-061269-1217. Bethesda, Md.: ERIC Document Reproduction Service, 1969, ED-029001.

Dixon, W. J. and Massey, F. J. Introduction to statistical analysis. (3rd. ed.) New York: McGraw-Hill, 1969.

Doob, L. Review of Pygmalion in the Classroom. The Key Reporter, Spring, 1969.

Draper, N. R. and Smith, H. Applied regression analysis. New York: Wiley, 1966.

Elashoff, J. D. Analysis of Covariance: A Delicate Instrument. *American Educational Research Journal* 6:3, 1969.

Elashoff, J. D. A model for quadratic outliers in linear regression. 1970. Submitted to *Journal of the American Statistical Association*.

Elashoff, R. M. Effects of errors in statistical assumptions. *International Encyclopedia of the Social Sciences*, 1968, 5, 132-142.

Hays, W. L. Statistics for psychologists. New York: Holt, Rinehart & Winston, 1963.

Huff, D. *How to lie with statistics*. New York: W. W. Norton, 1954.

Hutchins, R. Success in schools. *San Francisco Chronicle*, August 11, 1968, p. 2.

Kohn, H. Review of *Pygmalion in the Classroom*. *The New York Review of Books*, September 12, 1968, p. 31.

McCurdy, J. Testing of IQs in L. A. primary grades banned. *Los Angeles Times*, January 31, 1969.

Roberts, W. Voices in the classroom. *Saturday Review*, October 19, 1968, p. 72.

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

Rosenthal, R. Interpersonal expectations: Effects of the experimenter's hypothesis. In R. Rosenthal & R. Rosnow (Eds.) *Artifact in Behavioral Research*. New York: Academic Press, 1969 (a).

Rosenthal, R. Teacher expectation and pupil learning. Paper prepared for a conference on The Unstudied Curriculum sponsored by the Association for Supervision and Curriculum Development. Washington, D. C. January 8-11, 1969 (b).

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

Rosenthal, R. Another view of Pygmalion. Contemporary Psychology, 1970 (in press).

Rosenthal, R. & Jacobson, L. Teacher's expectancies: Determinants of pupils' IQ gains. Psychological Reports, 1966, 19, 115-118.

Rosenthal, R. & Jacobson, L. Teacher expectations for the disadvantaged. Scientific American, 1968, 218 (April), 19-23. (a)

†Rosenthal, R. & Jacobson, L. Pygmalion in the classroom: Teacher expectation and pupils' intellectual development. Copyright (c) 1968 by Holt, Rinehart and Winston, Inc., New York. Portions reprinted by permission of Holt, Rinehart and Winston, Inc. (b)

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

Scheffé, H. The analysis of variance. New York: Wiley, 1959.

Snow, R. E. Unfinished Pygmalion. Contemporary Psychology, 1969, 14, 197-200.

Snow, R. E. Still unfinished Pygmalion. Contemporary Psychology, 1970 (in press). The Urban Review, September, 1968.

Thorndike, R. L. Review of Pygmalion in the classroom. American Educational Research Journal, 1968, 5, 708-711.

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

Time, September 20, 1968, p. 62.

Tukey, J. W. Analyzing data: Sanctification or detective work? *American Psychologist*, 1969, 24(2), 83-91.

Walker, H. M. & Lev, J. *Statistical inference*. New York: Holt, Rinehart & Winston, 1953.

APPENDIX A: STATISTICAL TECHNIQUES

Analysis of Variance

Analysis of variance is a statistical technique designed to test the null hypothesis that the means of several groups are the same. A brief description of a standard two-way fixed effects analysis of variance with equal cell sizes will be used as an illustration. For a more general discussion of analysis of variance see the section on least squares. There are rc groups arranged in r rows and c columns; each group or cell contains the y scores of n individuals. For example, the c columns might be 2 treatments and the r rows might be 6 grades. Then we are interested in detecting differences between the means of the two treatment groups, differences between the means of the six grades, and interactions between treatments and grades.

To discuss the technique of analysis of variance it is helpful to write down a model for the individual scores, y_{ijk} , where i denotes rows, j denotes columns, and k denotes individuals within a group. Then the analysis of variance procedure rests on the assumptions that

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}$$

where the ε_{ijk} are independently and normally distributed with mean zero and variance σ^2 . The effects α_i , β_j , and γ_{ij} are defined sc that $\sum_i \alpha_i = 0$, $\sum_j \beta_j = 0$, $\sum_i \gamma_{ij} = 0$, $\sum_j \gamma_{ij} = 0$. In words, then, the observations in a particular cell (row i, column j for example) can be regarded as a random sample of n observations from a normal distribution with mean μ_{ij} and variance σ^2 . The observations in different cells are independent of each other but the variance in each cell is the same. We then wish to test the three null hypotheses: H_0 all $\alpha_i = 0$ or the means of the r rows are the same, H_0: all $\beta_j = 0$ or the means of the c columns are the same, H_0: all $\gamma_{ij} = 0$ or there are no differences in means between cells except those due to differences in row or column means.

The analysis of variance table is usually presented as follows:

Source	df	SS	MS
Rows	r-1	$cn\sum_i (\bar{x}_{i..}-\bar{x})^2$	$SS_R/(r-1)$
Columns	c-1	$rn\sum_j (\bar{x}_{.j.}-\bar{x})^2$	$SS_c/(c-1)$
Interaction	(r-1)(c-1)	$n\sum\sum_{ij} (\bar{x}_{ij.}-\bar{x}_{i..}-\bar{x}_{.j.}+\bar{x})^2$	$SS_I/(r-1)(c-1)$
Within cells	rc(n-1)	$\sum\sum\sum_{ijk} (x_{ijk}-\bar{x}_{ij.})^2$	$SS_{wc}/rc(n-1)$
Total	rcn-1	$\sum\sum\sum (x_{ijk}-\bar{x})^2$	

where $\bar{x}_{i..}$ for example denotes the mean of the observations in the i^{th} row.

To carry out the tests we note for example, that under the null hypothesis of equal row means

$$\frac{SS_R rc(n-1)}{SS_{wc}(r-1)}$$

is distributed as an F with r-1 and rc(n-1) degrees of freedom. The null hypothesis of equal row means is rejected at the α level of significance if F for rows is greater than the 95th percentile of the F distribution with r-1 and rc(n-1) degrees of freedom. (See for example Dixon and Massey (1969) or Hays (1963).)

This partition of the total sum of squares into mutually orthogonal (or independent) sums of squares due to each hypothesis is possible because the design is balanced (that is the sample size in each cell is equal).

Least Squares Procedure for Analysis of Variance

The section on analysis of variance shows the general formulas for a two-way fixed effects analysis of variance with equal cell sizes. When cell sizes are unequal the formulas are not so simple to write down and the sums of squares for rows, columns and interaction may not be orthogonal. To compute each particular analysis of variance we must fall back on the general principle underlying the derivation of the formulas, the least squares principle.

The model for an A x B classification, where the levels of A are denoted by $i = 1, 2, \ldots, r$ and the levels of B denoted by $j = 1, 2, \ldots, c$ is

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}$$

where there are n_{ij} observations in each cell and a total of N observations. The least squares principle states that the "best" estimates of μ, α_i, β_j, and γ_{ij} are those which minimize the sum of squared residuals about the line or those for which

$$\sum_{ijk}\sum\sum w_{ij} \, (y_{ijk} - \mu - \alpha_i - \beta_j - \gamma_{ij})^2$$

is minimized where the w_{ij} are some arbitrary system of weights. To derive these estimators, we must obtain the normal equations. The normal equation for μ is obtained by differentiating the sum of squared residuals with respect to μ and setting the result equal to zero. Thus the first normal equation is

$$0 = \sum_{ij}\sum w_{ij}n_{ij}\bar{y}_{ij} - N\mu\sum_{ij}\sum w_{ij} - \sum_{ij}\sum n_{ij}w_{ij}\alpha_i - \sum_{ij}\sum n_{ij}w_{ij}\beta_j - \sum\sum n_{ij}w_{ij}\gamma_{ij}$$

and there are r equations based on the α_i , c equations based on the β_j and rc equations based on the γ_{ij} . Usually when cell variances are equal we assume equal weights and the equations are somewhat simplified. For now, let us assume all $w_{ij} = 1$. The first equation then becomes

$$0 = \sum_{ij}\sum n_{ij}\bar{y}_{ij} - N\mu - \sum_i \alpha_i \sum_j n_{ij} - \sum_j \beta_j \sum_i n_{ij} - \sum_{ij}\sum n_{ij}\gamma_{ij} .$$

We notice, however, that our model for the cell means contains $1 + r + c + rc$ parameters and there are only rc cells and therefore only rc parameters can be estimated. So we must impose conditions on the parameters. These conditions can be identified as follows:

1) Select a set of wieghts corresponding to the levels of A, $\{u_i\}$ where $u_i \geq 0$ and $\Sigma u_i = 1$, and a set of weights corresponding to the levels of B, $\{w_i\}$ where $w_i \geq 0$ and $\Sigma w_i = 1$.

2) Then impose conditions

$$\sum_i u_i \alpha_i = 0$$

$$\sum_j w_j \beta_j = 0$$

$$\sum_i u_i \gamma_{ij} = 0 \quad \text{all } j \qquad \sum_j w_j \gamma_{ij} = 0 \quad \text{all } i .$$

With these conditions, the mean of the i^{th} level of A is $A_i = \sum_j w_j \mu_{ij}$, the mean of the j^{th} level of B is $B_j = \sum_i u_i \mu_{ij}$, and we define $\mu = \Sigma\Sigma\, u_i w_j \mu_{ij}$, and $\gamma_{ij} = \mu_{ij} - B_j - A_i + \mu$.

If, in fact, $\gamma_{ij} = 0$ for all i, j (no interaction), then the choice of weights will not affect SS_A or SS_B or any contrast among the α_i or β_j . Therefore, if there is no interaction, it will not matter what weights are chosen; the standard procedure would be to choose equal weights. If there is an interaction, the test of SS_{AB} is unaffected by the choice of weights but the main effects and tests on SS_A and SS_B will depend on the weights chosen.

If cell sizes are nearly equal and no other considerations suggest the use of unequal weights, the weights are usually chosen to be equal and the side conditions become

$$\sum_i \alpha_i = 0$$

$$\sum_j \beta_j = 0$$

$$\sum_i \gamma_{ij} = 0 \qquad \sum_j \gamma_{ij} = 0 .$$

Notice then that if equal weights are used and all cell sizes were equal that first normal equation becomes

$$0 = \sum_{ijk}\sum\sum y_{ijk} - N\mu$$

and the equations are quite simple. Otherwise the exact equations obtained will depend on the n_{ij} .

The F test for the null hypothesis that all $\alpha_i = 0$ when the β_j and γ_{ij} are included in the model is

$$\frac{SS_A/[r-1]}{SS_E/[N-rc]} \text{ where } SS_E = \sum_{ijk}\sum\sum (y_{ijk} - \hat{\mu} - \hat{\alpha}_i - \hat{\beta}_j - \hat{\gamma}_{ij})^2$$

where the $\hat{\mu}_1$, $\hat{\alpha}_i$, $\hat{\beta}_j$, and $\hat{\gamma}_{ij}$ are obtained by solving the normal equations and $SS_A = \sum_{ijk}\sum\sum (y_{ijk} - \hat{\mu}^1 - \hat{\beta}^1_j - \hat{\gamma}^1_{ij})^2$ where $\hat{\mu}^1$, $\hat{\beta}^1_j$, and $\hat{\gamma}_{ij}$ are obtained by solving the normal equations with all $\alpha_i = 0$. When the n_{ij} are all equal the estimators obtained under the two different conditions will be the same but when the n_{ij} are unequal $\hat{\mu} \neq \hat{\mu}^1$, etc.

For a full discussion, see Scheffé (1959). It should be noted that if there are any empty cells certain of the parameters will not be estimable.

Unweighted Means Analysis

Unweighted means analysis is a quick approximate method of calculating an analysis of variance with unequal cell sizes. The only justification for its use is the difficulty of calculating a full least squares analysis by hand. When the computer is available, the use of unweighted means analysis is not justified. The computations can be performed using the formulas shown in the section on analysis of variance except that $\bar{x}_{i..}$ is not the mean of all the observations but is now defined as $\bar{x}^1_{i..} = \sum_j \frac{x_{ij}}{c}$ and $\bar{x}^1_{.j.} = \sum_i \frac{\bar{x}_{ij}}{r}$, n is replaced by $n_h = \frac{rc}{\Sigma\Sigma \frac{1}{n_{ij}}}$, and the degrees of freedom within cells and total are replaced by N-rc and N-1 respectively where $N = \Sigma\Sigma n_{ij}$. See Winer (1962).

Example of the Effect of Using Proportional Weights

Refer to the discussion under least squares. An example will show what happens to the sums of squares for A and the sums of squares for B when we use unweighted means analysis, least squares with equal weights, and least squares with proportional weights (choosing $u_1 = u_2 = 1/2$, $w_1 = 5/6$, $w_2 = 1/6$). For a particular case where

Cell Sizes n_{ij}

	B_1	B_2
A_1	10	2
A_2	10	2

Cell Means $\bar{x}_{ij}$

	B_1	B_2
A_1	10	22
A_2	10	10

Unweighted means	$SS_A = 120$	$SS_B = 120$
Least squares with equal weights	$SS_A = 120$	$SS_B = 120$
Least squares with proportional weights	$SS_A = 24$	$SS_B = 120$

Thus, in estimating the effect of A, the cell with a mean of 22 receives much less weight when we take account of its small sample size by using proportional weights. The conclusion about B is unaffected by the use of proportional weights. Unweighted means and unweighted least squares give the same results; they would not if cell sizes were not exactly but only approximately proportional.

Analysis of Covariance

Analysis of covariance is an analysis of variance technique for situations in which information on a covariate x, a pretest or ability measure, etc. which is strongly predictive of the y observations is available. Thus it is used to test the null hypothesis that the means of several groups are the same based on the y scores after "adjustment" using the x scores. The covariance procedure reduces possible bias in treatment comparisons due to differences in the covariate x and increases precision in the treatment comparisons by reducing variability in the y scores "due to" variability in the covariate x.

The statistical model for a one-way analysis of covariance is composed of the four independent terms

$$y_{ij} = \mu + \alpha_i + \beta(x_{ij} - \bar{x}) + e_{ij} .$$

The e_{ij} are assumed to be an independent random sample from a normal distribution with mean zero and variance σ_e^2. The basic difference between analysis of variance and analysis of covariance is that in analysis of covariance the within cell variation ε_{ij} is divided into two parts, variability predicted by a linear regression on x, and unexplained variability e_{ij}.

The assumptions underlying the use of the analysis of covariance for testing the null hypothesis that all $\alpha_j = 0$ or there is no difference in group means for y not predictable from differences in group means for x are:

a) random assignment of individuals to groups,

b) y scores have a linear regression on x scores within each group,

c) the slope of the regression line is the same for each group

d) for individuals in the same group with the same x score, the y scores have a normal distribution,

e) the variance of the y scores among individuals with the same x score in the same group is the same for all x scores and all groups,

f) y scores can be represented by a linear combination of independent components: an overall mean, a group effect, a linear regression on x, and an error term.

For the details of the computations, see Dixon and Massey (1969). For a discussion of the importance of the assumptions see J. D. Elashoff (1969).

Simple Linear Regression

The technique of simple linear regression is based on the model that

$$y_i = \mu + \beta(x_i - \overline{x}) + \varepsilon_i$$

where the ε_i are independent and normally distributed with mean zero and variance σ^2 . The least squares estimators of μ and β are

$$\hat{\mu} = \overline{y}$$

$$\hat{\beta} = \frac{\Sigma(y_i - \overline{y})(x_i - \overline{x})}{\Sigma(x_i - \overline{x})^2} .$$

The model can arise in the situation when the x's are considered fixed and y is assumed to have a conditional normal distribution with mean $\mu + \beta(x_i - \overline{x})$ and variance σ^2 , or in the situation where x and y are assumed to have a bivariate normal distribution.

A test of whether two independent regression lines are parallel or have the slope β when the sample sizes n_1 and n_2 are equal and $\sigma_1^2 = \sigma_2^2$ is given by:

$$t = \frac{\hat{\beta}_1 - \hat{\beta}_2}{\frac{s_p}{\sqrt{n-1}}\sqrt{\frac{1}{s_{x_1}^2} + \frac{1}{s_{x_2}^2}}}$$

where $s_{x_i}^2 = \frac{\Sigma(x_{il} - \overline{x}_i)^2}{n-1}$ is the variance of the x's in sample i and

$$s_p^2 = \frac{s_{y_1 \cdot x}^2 + s_{y_2 \cdot x}^2}{2}$$ where

$$s_{y_1 \cdot x}^2 = (\frac{n-1}{n-2})(s_{y_i}^2 - \hat{\beta}_i^2 \, s_{x_i}^2)$$. The null hypothesis that

$\beta_1 = \beta_2$ is rejected at level α if $|t| > t_{2(n-2),\ 1-\alpha/2}$ or the $(1-\alpha/2)$ 100% of the t distribution with 2(n-2) degrees of freedom. See, for example, Dixon and Massey (1969) for a more complete discussion and the modification of the formulas for $n_1 \neq n_2$.

Correlation

The sample correlation between two variables y and x is given by

$$r = \frac{\Sigma(y_i - \bar{y})(x_i - \bar{x})}{\sqrt{\Sigma(x_i - \bar{x})^2 \ (y_i - \bar{y})^2}} .$$

When x and y have a bivariate normal distribution r is an estimate of ρ the population correlation between x and y. A test of the null hypothesis H_0: $\rho = 0$ is given by

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}.$$

Reject H_0 at level α if $|t| > t_{1-\alpha/2}$ with n-2 degrees of freedom.

When x is fixed and interest lies in the regression of y on x, r is mainly useful as a measure of the degree of fit of the regression line. The value of r^2 indicates the proportion of variance in the y variable predicted by the linear regression on x. If we denote the variance around the regression line as $s^2_{y \cdot x}$, then the "predicted variance" is $s^2_y - s^2_{y \cdot x}$ and

$$r^2 = \frac{s^2_y - s^2_{y \cdot x}}{s^2_y} .$$

Stepwise Regression

Stepwise linear regression is an ad hoc multiple linear regression technique in which predictor variables are entered one at a time into the equation in an attempt to obtain the "best" set of predictors. The basic procedure is as follows, at step one, the correlation with the dependent variable y of each of the possible predictor variables $x_1, \ldots, x_p$ is computed. Then the variable $x_{(1)}$ with the highest correlation with y is "entered first" and the regression of y on $x_{(1)}$ is computed. Then the partial correlations of the remaining x variables with y adjusted for $x_{(1)}$ are computed. The variable $x_{(2)}$ with the highest partial correlation with y is entered into the regression equation next. At each step, the x variable with the highest partial correlation with y adjusted for the x's already in the equation is entered. At each step then the x variable which will increase the multiple correlation coefficient R the most is entered. The square of the multiple correlation coefficient, R^2, gives the fraction of the variance of y which is "explained by" or predicted by the linear regression on the x variables. This basic procedure called "forward selection" is modified in two ways in a standard stepwise regression program such as BMD 02R. At each stage, and for each x variable not in the equation an F-statistic is calculated to allow determination of the statistical significance of the partial correlation of x with y adjusted for the x's in this equation. If the F-statistic for the x with the highest partial correlation is not larger than a prespecified critical value of F, the procedure is terminated and no new variables are entered into the equation. In addition at each stage, for each x variable in the equation, an F-statistic is computed based on the partial correlation of x with y adjusted for the other x variables in the equation; if this F value falls below a prespecified F-to-remove value that x variable is deleted from the equation. That is at each stage we check back to make sure that all the variables in the equation still make a reasonable contribution to R^2 (Draper and Smith (1966) provide a useful introduction to multiple regression and stepwise regression.)

The BMD 02R program offers an additional modification to the general stepwise regression procedure. Any of the variables may be forced to enter the equation first irrespective of the value of their correlation with y . Additional x variables may be forced into the equation in a predetermined or partially predetermined order. That is, if two variables are designated to be forced in at level j , the variable with the highest partial correlation will be entered first and the other variable entered next; then the program proceeds to the next level of forced variables.

Clearly then, leaving all the variables free, stepwise regression provides an ad hoc procedure for determining the relative importance of the x variables as predictors of y and for obtaining the "best" set of predictors. There is of course no guarantee that the variables selected will constitute the "best" set. Using the option of forcing variables in, we may assess the predictive power of a variable by itself versus its additional predictive power after other variables have been included.

Test Scores and Norms

The primary outcome of a test administration is a raw score, usually a number indicating how many items in a test or part an individual answered correctly. As it stands, this number is useful for research purposes and it should always be retained in whatever records are kept about this test performance. For many practical purposes, however, the raw score must be transformed in some way or related to other information to be interpreted properly.

Norms are tables of score distributions obtained in various reference groups. They relate raw score scales to proposed conversion scores, like mental age, IQ, or grade equivalents. Most test manuals will provide norms, at least for a "national" sample of people for whom the test is presumed appropriate. The best manuals, however, contain carefully specified breakdowns of norm tables to show distributions for sex, grade, geographic or social strata, or other subgroups of importance.

With norms and a standard error of measurement in hand, it is possible to interpret scores more completely. A child whose IQ score has changed

10 points in the past year may not be considered unusual if it is seen that 10 IQ points equals 4 raw score points at this part of the test range and the raw score standard error is 5. For another child elsewhere in the range, a 10 point IQ change might be considered substantial. One cannot tell without knowing raw score equivalents and standard errors.

Often, published norms are not complete or are extrapolated beyond the range of the distributions available in norm samples. Use of such extrapolations, whether computed by test maker or user, cannot be recommended. The central question in using any particular score or norm conversion is whether the obtained scale of measurement is meaningful for the particular population and interpretation intended.

Reliability

The reliability of a variable X, such as scores on an IQ test, is an estimate of the test's accuracy as a measuring instrment. Reliability can be defined in different ways depending on the model we choose to represent variation in obtained X scores. In practical situations it may be difficult to estimate reliability and many different formulas have been advanced, some based on correlations between equivalent forms of the test, some on measures of internal consistency of the test, and some on correlations showing the stability of the obtained score over repetitions of the test.

A standard model proposes that the observed score X is a combination of a true score x and an error e, that is

$$X = x + e$$

where x and e are independent and $\mu_e = 0$. Then the reliability of

X is defined as the ratio of the true variance to observed variance, or the proportion of variance in X not due to error

$$R_x = \frac{\sigma_x^2}{\sigma_X^2} = \frac{\sigma_x^2}{\sigma_x^2 + \sigma_e^2} \ .$$

If x remains constant and X is measured twice then the correlation between X_1 and X_2 is R_x .

The Binomial Distribution

Suppose there are n independent experiments (or items) which can each result in a success or failure (right or wrong) and that in each experiment the probability of a success is p . Then the probability distribution of the number of successes in n trials, X , is the binomial distribution and

$$P(X = x) = \binom{n}{x} p^x (1 - p)^{n-x} \quad \text{for} \quad x = 0, 1, \ldots, n$$

and $\binom{n}{x} = \frac{n!}{x!(n-x)!}$. (See, for example, Hays (1963)). The expected number or mean number of successes in n trials or items is np and the variance is np(1-p).

Sign Test

The sign test is used for testing hypotheses about the median of a population or the median difference between matched pairs. To test the null hypothesis that all the observations (or for matched pairs all the differences) come from populations with median zero the observations are classified merely as positive or negative and the null hypothesis that the common median is zero is rejected if the number of positive signs is too large or too small.

Under the assumptions that the n observations are independent of each other and there are no zeros (scores which are neither positive nor negative) and the null hypothesis that the median is zero, the probability of a positive score is one-half and the number of positive scores, r , has a binomial distribution with parameters n and $p = 1/2$. If there are only a few zeros the sample size is reduced and the test carried out on the nonzero observations. See, for example, Dixon and Massey (1969) for a description of the test and tables for its use. If the sign test is to be used for matched pairs we must assume in addition, random assignment to treatments within pairs and each member of the pair treated the same except for the treatment.

Expected Number Correctly Classified

In a particular group there are n children, c of whom are in the control group, and t of whom are in the experimental group. If we randomly select t of the n children what is the expected number of experimental children, e , in the t children selected? Under the null hypothesis that the treatment does not affect posttest or gain scores, selection of the t children on the basis of posttest or gain scores should be equivalent to selection at random with respect to the two treatment groups.

The number of experimental children selected among the t will have a hypergeometric distribution with parameters n, t, c .

$$P\ (n_t = e) = \frac{\binom{t}{e}\binom{n-t}{t-e}}{\binom{n}{t}}$$

The mean of this distribution or the expected value of e is

$$E\ (e) = \frac{t^2}{n}$$

and $$\text{Var}\ (e) = \frac{t^2\ (n-t)^2}{n^2\ (n-1)} .$$

See, for example, Hays (1963).

Therefore in group i , we expect to classify correctly t_i^2/n_i children by chance; since the groups are independent, the expected number correctly classified across all the groups is $\Sigma\ t_i^2/n_i$ and the variance is $\Sigma\ \dfrac{t_i^2\ (n_i - t_i)^2}{n_i^2\ (n_i - 1)}$.

Wilcoxon Rank Sum Test

The Wilcoxon rank sum test (also referred to as the Mann Whitney U) is a test of the null hypothesis that two samples both represent a random sample from the same population. It is sensitive to shifts in location and thus is frequently used as a test of whether two samples come from populations with the same mean or median assuming that the distributions of the two populations are the same in other respects.

The two samples are pooled and all the observations are rank ordered. Then the observations are replaced by their ranks and the sum of the ranks for one sample is computed. If the sum of the ranks is too large or too small we reject the null hypothesis that the two samples are drawn at random from identical populations. Tables of the distribution of the rank sum are available in such books as Dixon and Massey (1969).

The assumptions underlying the use of this test are that observations are continuous and therefore no tied ranks occur and that each sample constitutes a random sample from one population. (Procedures for applying the test when some ties occur have been developed.)

Wilcoxon Signed Rank Test

The Wilcoxon signed rank test is used for testing hypotheses about the mean or median of a population (or the mean or median difference between matched pairs.) To test the null hypothesis that the observations are drawn from a population with a mean of zero, the observations are ranked from smallest to largest in absolute value. Then the sum of the ranks of the positive observations is computed. The null hypothesis is rejected if the sum of the positive ranks is too small or too large, see Dixon and Massey (1969) for tables.

The signed rank test is based on the assumptions that the observations are continuous (there are no ties) and there are no zeros (all observations are either positive or negative.) Procedures exist :or performing the test when zeros or ties exist. It must be further assumed that all observations come from symmetric populations with a common median.

APPENDIX B: LISTING OF THE DATA SUPPLIED BY ROSENTHAL AND JACOBSON

The cards are listed in order by grade, track, experimental group, sex and minority group. The codes used on the cards are:

G = Grade

A = Ability track

- 1 = slow
- 2 = medium
- 3 = fast

T = Treatment group

- 0 = Control
- 1 = Experimental

M = Minority group

- 0 = Non-Mexican
- 1 = Mexican

S = Sex

- 0 = Female
- 1 = Male

ID#	G	A	T	M	S	Age	Total IQ				Verbal IQ				Reasoning IQ			
							1	2	3	4	1	2	3	4	1	2	3	4
31	1	1	0	0	0	5.57	61	82	87	81	79	87	97	88	18	71	72	73
47	1	1	0	0	0	5.95	84	86	86	82	104	97	92	84	17	71	78	81
57	1	1	0	0	0	6.28	75	62	85	75	83	75	85	80	65	40	85	71
77	1	1	0	0	0	6.20	45	58	76	75	65	77	96	81	0	0	39	68
12	1	1	0	1	0	6.42	79	90	82		97	96	89		31	85	75	
84	1	1	0	1	0	5.86	61	86	86	94	85	80	87	87	0	95	82	103
60	1	1	0	0	1	5.61	100	70	95	98	123	83	91	97	57	51	100	99
97	1	1	0	0	1	5.95	72	94	94	77	91	116	99	86	17	71	86	63
21	1	1	0	0	1	6.32	57	84	93	85	79	86	90	80	0	80	96	91
95	1	1	0	0	1	5.78	88	85	81	84	107	90	83	75	35	81	80	95
78	1	1	0	0	1	6.28	65	76	78	67	83	83	93	71	32	63	56	61
86	1	1	0	0	1	6.42	81	80			93	82			56	79		
36	1	1	0	0	1	5.82	93	85			120	96			17	68		
89	1	1	0	0	1	5.53	72	77			94	87			18	58		
100	1	1	0	0	1	5.57	79	111	134		97	112	131		36	106	135	
71	1	1	0	0	1	6.03	111	112	105		123	115	122		100	110	91	
22	1	1	0	0	1	5.82	86	94	101	89	103	96	97	88	34	92	113	90
48	1	1	0	0	1	6.20	85	87	96	87	103	102	103	92	52	64	92	81
25	1	1	0	1	1	5.95	67		96	95	89		83	81	0		124	117
63	1	1	1	0	0	5.70	60	94	97	89	70	104	103	80	49	82	90	98
70	1	1	1	1	1	5.53	61	110	106	104	65	100	106	100	58	124	107	111
98	1	2	0	0	0	5.82	108	102	103		117	106	113		100	96	97	
76	1	2	0	0	0	5.49	118	102	123	109	135	110	123	101	102	94	126	120
90	1	2	0	0	0	5.90	98	100	96		112	104	100		80	97	90	
56	1	2	0	0	0	5.70	91	97	90	101	95	110	96	102	82	82	81	98
35	1	2	0	0	0	6.15	85		78		107		92		16		57	
30	1	2	0	1	0	6.28	84	106	95	94	86	106	95	74	80	106	93	122
85	1	2	0	0	1	5.82	55	102	95	104	76	119	101	101	0	86	88	111
45	1	2	0	0	1	6.28	80	83	87		99	95	110		16	68	65	
66	1	2	0	0	1	5.61	100	124		99	111	111		93	84	142		110
3	1	2	0	0	1	5.61	100	108	120		137	123	130		18	96	97	
46	1	2	0	0	1	5.57	92	101	105		118	119	122		0	83	91	
88	1	2	0	0	1	5.95	39	103	104	94	54	121	101	74	0	88	106	123

ID#	G	A	T	M	S	Age	Total IQ				Verbal IQ				Reasoning IQ			
							1	2	3	4	1	2	3	4	1	2	3	4
44	1	2	0	0	1	5.70	91	113	96	113	109	116	122	102	49	110	70	127
94	1	2	0	0	1	5.65	94	111	113		106	130	120		64	98	105	
9	1	2	0	0	1	5.65	94	100	111	113	110	101	111	96	64	98	111	137
24	1	2	0	1	1	5.82	74	91	88	97	96	102	94	68	0	77	82	143
10	1	2	1	0	0	5.74	84	120	107	105	105	106	104	102	17	148	110	112
1	1	2	1	1	0	5.70	88	85	128	101	109	107	133	102	18	44	122	98
53	1	2	1	0	1	5.49	100	94	108		113	110	114		80	71	105	
93	1	2	1	0	1	5.95	116	104	137	129	138	106	144	138	101	103	128	123
74	1	3	0	[illegible]	0	5.95	94	151	115		101	144	115		87	178	118	
32	1	3	0	0	0	5.70	102	132	128		119	126	119		77	140	142	
16	1	3	0	0	0	5.49	117	112	126	109	128	114	126	116	102	110	126	101
23	1	3	0	0	0	6.28	91	86	96	109	110	89	102	101	57	83	93	122
64	1	3	0	0	0	6.24	90	104	93	87	96	107	94	98	83	101	91	75
73	1	3	0	0	0	6.24	106		91	109	112		102	112	96		80	105
55	1	3	0	0	0	5.82	112	106		103	119	114		125	103	99		79
39	1	3	0	0	0	6.11	101	100	84		105	103	93		98	94	76	
2	1	3	0	0	1	6.15	122	130			133	130			111	130		
68	1	3	0	0	1	6.32	111	114	126	144	122	110	122	149	104	123	130	142
54	1	3	0	0	1	6.03	100	133	125	100	103	128	114	92	96	142	142	112
65	1	3	0	0	1	6.36	108	117	125	117	126	127	129	125	97	105	121	107
72	1	3	0	0	1	5.74	117	125	131	124	139	139	141	145	98	109	122	104
75	1	3	0	0	1	6.45	130	147	119	174	183	166	221	168	102	133	94	195
43	1	3	0	0	1	6.20	116	108	108	131	139	130	119	136	100	93	97	123
5	1	3	0	0	1	5. 9	97	106	111	105	117	110	114	108	58	101	108	101
19	1	3	0	0	1	5.74	120	108	110		129	106	98		111	115	132	
11	1	3	0	0	1	6.07	86	101	103	119	99	104	113	117	59	95	109	121
41	1	3	0	0	1	5.86	92	84	102	110	90	86	101	103	96	83	108	118
51	1	3	1	0	0	6.42	95	96	113	97	100	97	108	103	90	93	120	90
38	1	3	1	0	0	6.11	98	97			101	94			92	100		
17	1	3	1	0	1	6.11	92	127			105	118			72	140		
173	2	1	0	0	0	6.74	79	85	76		89	94	78		53	73	75	
164	2	1	0	0	0	7.15	73	83	67	68	81	87	71	67	57	79	64	70

ID#	G	A	T	M	S	Age	Total IQ				Verbal IQ				Reasoning IQ			
							1	2	3	4	1	2	3	4	1	2	3	4
170	2	1	0	0	0	7.24	75	77	90	87	94	87	108	106	39	66	75	66
127	2	1	0	0	0	6.90	86	99	95	87	96	106	120	97	72	92	78	75
136	2	1	0	0	0	6.86	89	70	88	80	87	77	104	83	93	62	74	75
140	2	1	0	0	0	6.86	80	82	85		90	88	109		64	77	66	
166	2	1	0	0	0	7.53	80	78	98	88	82	84	111	97	77	71	87	77
157	2	1	0	1	0	7.07	85	93	97	90	85	85	95	83	85	111	102	99
162	2	1	0	1	0	6.86	93	88	102	95	87	92	89	80	102	82	121	118
176	2	1	0	1	0	6.82	78	81	84	88	78	77	88	91	76	88	79	83
169	2	1	0	0	1	7.20	64	76	84	78	50	86	100	97	75	66	71	55
167	2	1	0	0	1	6.74	89	94	84	92	89	104	89	96	86	89	78	86
171	2	1	0	1	1	7.86	59	67	81	76	64	62	79	65	52	73	84	91
161	2	1	0	1	1	7.86	88	90	97	97	79	75	90	85	104	138	107	119
160	2	1	1	0	0	6.78	97	89		84	94	86		70	100	91		102
158	2	1	1	0	1	7.53	80	80	91		92	94	111		66	68	77	
159	2	1	1	1	1	8.11	74	85	86		74	88	81		72	84	94	
131	2	1	1	1	1	6.82	91	93		78	100	99		75	82	91		80
174	2	1	1	1	1	6.61	82	89	101		91	82	92		67	96	113	
156	2	2	0	0	0	7.11	113	111	120		108	122	123		121	99	117	
133	2	2	0	0	0	6.49	100	112			92	120			108	103		
154	2	2	0	0	0	7.95	74	64	70	61	88	74	74	56	52	48	67	65
143	2	2	0	1	0	7.11	84	82	104	99	97	89	95	88	70	75	117	115
137	2	2	0	0	1	7.28	84	94	101	93	91	101	107	87	77	88	93	100
126	2	2	0	0	1	7.36	83	92	100	81	92	100	103	75	73	85	98	89
130	2	2	0	0	1	6.78	111	115	125	123	109	119	129	134	114	110	122	111
120	2	2	0	0	1	7.03	35	104	110		110	106	125		58	100	96	
147	2	2	0	0	1	6.45	95	98	107	112	105	115	119	116	84	87	94	106
165	2	2	0	0	1	6.90	77	61	86	90	87	79	84	88	59	0	89	94
134	2	2	0	0	1	6.57	97	95	92	95	104	119	118	98	91	80	77	94
172	2	2	0	0	1	7.32	94	80	87		112	96	114		79	65	70	
146	2	2	0	0	1	7.36	79	83			82	87			76	77		
163	2	2	0	1	1	7.24	76	73	84	72	77	78	84	66	75	66	83	80

ID#	G	A	T	M	S	Age	Total IQ				Verbal IQ				Reasoning IQ			
							1	2	3	4	1	2	3	4	1	2	3	4
145	2	2	0	1	1	6.90	91	90	101	104	100	85	104	110	81	98	97	97
132	2	2	0	1	1	6.65	83	98	98	113	90	101	97	116	71	96	101	107
139	2	2	1	0	0	6.95	108	92	132	123	106	97	119	123	111	89	208	125
149	2	2	1	0	0	6.82	89	92	90	98	97	93	90	83	82	91	90	119
135	2	2	1	0	1	6.49	121	96	107	112	108	115	134	115	114	81	88	106
125	2	3	0	0	0	7.42	119	114	109	127	120	124	125	139	116	106	97	119
110	2	3	0	0	0	6.74	86		114	108	89		99	103	83		152	115
104	2	3	0	0	0	6.78	114	94	99	95	131	128	135	115	97	75	77	73
112	2	3	0	0	0	6.95	99	89	101		111	105	112		89	76	88	
113	2	3	0	0	0	6.95	99	102	106		111	125	119		92	87	93	
115	2	3	0	0	0	6.90	100	114	104	106	100	114	97	101	101	114	113	113
151	2	3	0	0	0	6.70	103	114	94	104	101	104	88	109	104	129	106	100
155	2	3	0	0	0	7.32	93	85	82	87	96	88	82	81	87	80	82	96
107	2	3	0	1	0	7.24	97	88	100	127	97	109	121	142	97	76	83	113
122	2	3	0	0	1	7.35	112	110	114	119	112	111	103	104	112	107	141	152
148	2	3	0	0	1	6.78	115	123	118	123	109	119	122	119	127	128	114	127
128	2	3	0	0	1	7.45	101	106	109		99	101	102		103	110	118	
121	2	3	0	0	1	6.65	96	107	107	121	123	130	137	129	71	90	86	113
118	2	3	0	0	1	7.42	94	89	113	105	111	106	106	95	84	77	119	125
105	2	3	0	0	1	6.49	126	120	130	119	126	124	127	129	126	115	134	109
103	2	3	0	0	1	6.49	129	117	127	154	133	147	119	167	126	95	134	146
141	2	3	0	0	1	6.70	101	111	119	113	110	121	130	126	93	100	112	103
123	2	3	0	0	1	7.11	115	118	123	144	121	122	123	144	108	114	123	144
116	2	3	0	0	1	7.24	110	123	121	127	119	126	108	127	102	120	200	127
117	2	3	1	0	0	6.57	105	113	125	122	107	119	118	130	104	106	132	114
119	2	3	1	0	0	6.82	109	104	134		103	115	121		113	93	211	
111	2	3	1	0	0	7.36	101	111	114		111	125	120		90	102	106	
102	2	3	1	0	1	7.15	123	113	117	135	115	102	109	129	133	128	123	143
109	2	3	1	0	1	6.86	109	122	123	126	117	118	121	148	102	126	127	110
106	2	3	1	0	1	7.15	133	142	202	131	133	151	202	114	133	128	202	180

ID#	G	A	T	M	S	Age	Total IQ				Verbal IQ				Reasoning IQ			
							1	2	3	4	1	2	3	4	1	2	3	4
236	3	1	0	0	0	7.49	95	83	98	91	104	89	98	74	85	75	98	110
238	3	1	0	0	0	7.49	112	96			123	86			100	109		
232	3	1	0	1	0	9.15	103	78	88	86	103	71	82	63	101	88	96	117
252	3	1	0	1	0	8.03	85	89	86	77	103	95	92	74	56	80	81	81
242	3	1	0	1	0	8.53	95	79	85	77	94	74	77	65	97	87	97	93
244	3	1	0	0	1	7.78	64	80	74	76	68	95	69	83	58	59	80	68
237	3	1	0	0	1	7.70	117	103		93	126	112		89	108	90		96
251	3	1	0	0	1	7.65	127	106		95	145	94		95	108	120		95
229	3	1	0	0	1	7.86	107	103	104	99	120	104	106	88	93	101	100	112
226	3	1	0	0	1	8.53	104	88	105		91	90	102		122	87	109	
241	3	1	0	0	1	7.61	97	107	113	99	113	117	116	109	80	97	107	88
245	3	1	0	0	1	8.32	94	93	109	89	113	105	132	91	73	81	92	87
246	3	1	0	0	1	7.86	107	108	111	99	127	117	123	97	84	97	100	101
230	3	1	0	0	1	7.53	98	116	110	97	110	127	122	96	85	105	97	98
235	3	1	0	1	1	8.20	109	106	105	94	115	123	118	112	101	90	93	75
247	3	1	0	1	1	7.57	91	80	93	79	90	85	93	78	92	74	93	80
228	3	1	1	0	0	7.74	103	93	88	79	125	87	86	68	79	102	92	92
248	3	1	1	0	0	7.53	86	90	88	86	100	83	80	83	66	101	101	89
240	3	1	1	0	1	8.36	93	83	92	89	103	99	95	94	84	68	89	83
227	3	1	1	0	1	7.45	115	108	109		126	113	118		101	102	98	
249	3	1	1	0	1	8.36	110		111		133		125		87		98	
206	3	2	0	0	0	7.70	104	90	106	106	95	90	111	103	116	90	99	108
221	3	2	0	0	0	7.99	94	94	99		91	96	99		100	92	99	
205	3	2	0	0	0	8.20	99	88	87	80	91	94	100	88	109	82	72	73
231	3	2	0	0	0	7.61	102	86	85	93	109	104	100	97	96	66	67	88
213	3	2	0	0	0	8.20	78	80	79	87	74	85	82	94	80	74	76	78
215	3	2	0	0	0	7.82	95	104			114	111			74	94		
224	3	2	0	1	0	8.24	107	120	108	122	114	122	120	128	97	117	96	116
239	3	2	0	0	1	7.95	113	100	109	103	122	113	131	90	104	85	93	123
204	3	2	0	0	1	7.95	104	103	112	103	118	136	145	127	88	74	89	82
214	3	2	0	0	1	7.74	121	113	105	118	129	139	114	145	111	89	95	96
217	3	2	0	0	1	7.57	111	108	100	109	124	108	121	123	96	108	77	98

ID#	G	A	T	M	S	Age	Total IQ 1	Total IQ 2	Total IQ 3	Total IQ 4	Verbal IQ 1	Verbal IQ 2	Verbal IQ 3	Verbal IQ 4	Reasoning IQ 1	Reasoning IQ 2	Reasoning IQ 3	Reasoning IQ 4
212	3	2	0	0	1	7.70	105	108	103		101	130	115		112	87	92	
202	3	2	0	0	1	8.03	110	111	136	140	117	108	123	122	100	115	156	174
219	3	2	0	0	1	8.15	92	102	96	85	92	98	94	85	92	110	97	85
191	3	2	0	0	1	7.86	103	111			120	137			84	88		
208	3	2	0	1	1	8.45	78	84	85	81	78	80	77	71	78	91	94	90
203	3	2	0	1	1	8.24	121	138	127	165	142	146	141	161	104	131	113	170
223	3	2	1	1	1	7.86	88	120	97	97	95	122	104	104	78	117	90	88
194	3	3	0	0	0	8.32	101	106	104		103	116	119		100	96	89	
182	3	3	0	0	0	7.82	95	99	104	102	96	111	124	101	93	86	85	105
190	3	3	0	0	0	7.78	104	115	126	118	107	111	118	110	103	118	140	130
193	3	3	0	0	0	8.28	117	114	126	135	111	112	117	115	126	116	140	169
186	3	3	0	0	0	8.15	109	108	119		102	113	109		119	101	134	
192	3	3	0	0	0	8.49	108	111	142	137	118	128	154	141	98	97	130	134
225	3	3	0	0	0	7.99	71	85		96	80	99		100	56	70		90
177	3	3	0	0	0	7.99	85	109	96	99	76	99	89	87	94	120	102	115
207	3	3	0	1	0	7.78	85	92	105		90	95	101		78	89	110	
195	3	3	0	0	1	7.95	108	114	141		101	121	137		116	107	145	
187	3	3	0	0	1	8.07	103	114	120	103	110	134	129	102	93	98	110	104
185	3	3	0	0	1	7.86	125	120	132		141	192	300		109	86	100	
222	3	3	0	0	1	7.99	76	102		96	69	106		98	83	96		92
218	3	3	0	1	1	7.65	90	101	86	92	80	90	87	95	105	111	84	88
220	3	3	1	0	0	8.32	79	86	86	72	73	89	84	79	88	81	89	64
197	3	3	1	0	0	7.82	106	108	129	117	106	111	126	113	106	105	133	121
211	3	3	1	0	0	8.28	104	103	108	95	100	112	105	102	107	93	112	87
184	3	3	1	0	0	7.61	106	118	107	114	117	141	127	124	96	100	87	104
209	3	3	1	0	1	7.53	112	110	120	108	133	115	137	129	88	105	104	89
216	3	3	1	0	1	7.70	105	114	125	121	101	110	111	108	112	119	149	140
189	3	3	1	0	1	8.32	96	102	104		96	108	107		96	96	99	
180	3	3	1	0	1	7.95	94	110	103		112	113	116		77	107	89	
179	3	3	1	0	1	7.74	123	127			115	130			134	124		
324	4	1	0	0	0	9.42	68	72	74	82	70	74	77	71	65	69	70	94
298	4	1	0	0	0	8.45	89	101			89	110			89	91		
328	4	1	0	0	0	9.36	101	97	103	97	104	104	100	81	98	89	107	130

ID#	G	A	T	M	S	Age	Total IQ				Verbal IQ				Reasoning IQ			
							1	2	3	4	1	2	3	4	1	2	3	4
320	4	1	0	0	0	8.53	115	102	112	77	117	121	129	93	114	82	97	60
316	4	1	0	1	0	8.86	84	94	103	104	85	90	101	90	82	102	105	125
318	4	1	0	1	0	9.15	80	74	73	75	70	81	79	61	94	65	65	89
314	4	1	0	0	1	9.45	68	64	63	60	68	63	72	64	68	65	53	55
321	4	1	0	0	1	9.15	72		57	73	82		74	80	60		0	67
317	4	1	0	0	1	9.11	98	100			101	126			94	82		
326	4	1	0	0	1	8.86	102	103		109	117	129		120	85	84		99
309	4	1	0	0	1	8.53	90	96	85		88	93	84		94	97	87	
313	4	1	0	0	1	9.32	94	84	92	68	101	104	108	86	86	64	78	49
315	4	1	0	0	1	8.99	100	106	102	97	105	104	100	95	96	108	104	98
319	4	1	0	0	1	8.70	85	98	97	107	84	92	95	91	86	107	100	132
312	4	1	0	0	1	8.78	105	104	106	100	114	130	133	103	95	85	88	97
305	4	1	0	0	1	9.32	117	109	110	104	151	164	141	112	95	83	89	95
329	4	1	0	1	1	9.45	89	97	102	109	88	96	104	90	91	99	100	152
310	4	1	0	1	1	8.70	101		75	100	95		86	91	106		63	111
330	4	1	1	1	0	9.11	89	80	85	82	88	82	91	78	91	77	79	88
311	4	1	1	1	0	9.20	90	90	102	93	76	79	95	80	109	105	115	117
325	4	1	1	0	1	8.53	86	100	100	99	86	90	102	89	86	113	97	117
322	4	1	1	0	1	9.24	90	91	100	88	100	101	108	106	79	81	90	71
303	4	2	0	0	0	9.15	101	99	102	83	101	106	102	80	101	91	102	87
299	4	2	0	0	0	8.99	96	92	97		99	113	117		92	72	80	
290	4	2	0	0	0	8.82	94	97	100	99	94	105	106	106	94	87	94	92
286	4	2	0	0	0	9.24	100	95	93	89	102	110	102	96	96	81	84	83
287	4	2	0	0	0	9.15	103	91	93	96	109	106	115	94	94	74	72	96
304	4	2	0	1	0	9.28	97	89	88	80	93	89	87	72	105	89	89	88
307	4	2	0	1	0	9.42	77	80	90	76	74	72	88	70	80	91	93	82
291	4	2	0	1	0	8.74	114	101	112	89	114	100	114	81	114	103	107	97
297	4	2	0	0	1	8.70	113	125	112	113	115	156	151	110	111	104	89	118
293	4	2	0	0	1	9.28	81	112	111	95	75	110	108	86	89	118	114	109
300	4	2	0	0	1	8.70	117	111	110	118	134	131	134	127	102	95	92	111
292	4	2	0	0	1	9.28	106	120	114	98	112	131	114	109	99	112	114	88

ID#	G	A	T	M	S	Age	Total IQ				Verbal IQ				Reasoning IQ			
							1	2	3	4	1	2	3	4	1	2	3	4
306	4	2	0	0	1	8.86	125	97	105	93	132	116	125	109	117	77	90	78
294	4	2	0	0	1	8.99	105	100	104	82	92	95	104	79	123	108	104	85
258	4	2	0	0	1	8.95	106	96	90	100	116	113	105	98	96	78	75	101
255	4	2	0	0	1	8.99	121	118			121	127			123	108		
288	4	2	0	1	1	8.99	111	111	117	113	123	146	123	119	99	89	111	108
295	4	2	1	0	0	9.20	90	87	93	88	90	76	78	80	90	101	121	99
308	4	2	1	1	0	9.15	107	125	117	115	103	125	109	99	114	125	128	144
259	4	2	1	1	1	9.95	95	79	81		105	78	86		86	81	76	
283	4	3	0	0	0	9.07	110	110		109	115	107		139	107	114		90
265	4	3	0	0	0	8.86	121	107			109	109			139	105		
272	4	3	0	0	0	9.49	91	91	93	84	105	95	106	89	74	85	82	78
273	4	3	0	0	0	9.03	132	103	117	107	123	100	111	100	144	107	123	115
276	4	3	0	0	0	9.28	105	117	116		152	142	160		75	101	94	
274	4	3	0	0	0	9.28	120	103	114	112	117	101	106	98	126	105	126	136
268	4	3	0	0	0	8.74	119	118	122	126	134	116	114	118	105	124	133	138
280	4	3	0	0	0	8.70	120	101	121	117	120	125	145	127	120	80	103	107
253	4	3	0	0	0	8.99	123	108	119	112	130	108	141	107	116	108	104	119
271	4	3	0	0	0	9.32	135	114	119	142	151	123	137	146	126	104	108	136
257	4	3	0	0	0	9.42	108	108	125	109	110	129	157	104	106	91	107	114
261	4	3	0	0	0	10.32	97	99	125	98	108	106	125	103	86	91	125	93
269	4	3	0	0	0	8.74	149	118	139	138	149	116	150	122	149	124	126	162
281	4	3	0	0	0	9.28	120	101	142		126	97	137		112	105	160	
277	4	3	0	0	0	9.15	121	119	162	124	119	125	144	114	128	113	262	138
264	4	3	0	0	0	9.15	154	119	153	112	154	106	162	96	154	144	139	138
284	4	3	0	0	0	8.45	146	120	149	128	131	103	138	121	167	155	174	135
263	4	3	0	0	1	9.07	131	129	140	139	136	133	129	118	129	126	163	185
270	4	3	0	0	1	9.45	110	103	135	112	117	139	157	119	103	82	118	107
260	4	3	0	0	1	8.99	116	113	135	115	145	151	164	134	96	89	117	101
266	4	3	0	0	1	8.53	115	127			128	159			104	105		
262	4	3	0	0	1	8.61	138	108			164	112			121	105		

ID#	G	A	T	M	S	Age	Total IQ				Verbal IQ				Reasoning IQ			
							1	2	3	4	1	2	3	4	1	2	3	4
254	4	3	1	0	0	9.20	134	125	127	106	134	132	127	106	134	119	127	106
278	4	3	1	0	0	9.32	109	100	122	114	112	92	113	112	107	117	137	115
285	4	3	1	0	0	8.42	127	135	155	138	167	293	174	130	102	107	138	148
267	4	3	1	0	1	8.95	158	147	165	159	163	152	267	168	145	135	142	147
282	4	3	1	0	1	8.53	137	107	132	110	171	118	136	112	114	97	129	109
387	5	1	0	0	0	9.95	67	76	85	75	63	76	79	68	68	72	93	82
399	5	1	0	0	0	9.78	91	84	78	81	86	87	84	87	91	76	71	73
380	5	1	0	1	0	11.28	56	78	83	81	46	64	66	60	62	87	110	131
362	5	1	0	0	1	9.90	85	76	83	82	82	72	82	79	87	79	85	86
366	5	1	0	0	1	9.53	77	112		95	90	116		133	61	105		71
393	5	1	0	0	1	9.95	90	98		106	84	111		109	91	81		103
388	5	1	0	0	1	10.49	85	89	96		91	93	102		72	80	91	
372	5	1	0	0	1	10.11	103	95	120	103	109	106	132	112	98	84	110	95
401	5	1	0	0	1	10.82	83	108	114		90	110	124		70	96	107	
371	5	1	0	0	1	10.11	97	101	112	119	101	121	132	116	92	84	96	122
381	5	1	0	0	1	10.03	80	107	103	96	91	103	90	83	66	103	122	117
397	5	1	0	1	1	9.86	86	99	110	81	77	80	91	75	90	116	147	88
384	5	1	1	1	0	9.61	92	107	121	109	76	82	93	90	101	136	251	150
375	5	1	1	1	0	10.07	85	99	96	100	80	95	88	95	90	102	107	108
370	5	1	1	1	0	10.74	90	87	89	93	100	103	111	103	80	74	72	84
391	5	1	1	0	1	11.45	78	110	104	107	85	116	118	114	65	94	95	101
363	5	2	0	0	0	10.24	90	103		85	87	95	88	97	95	101		75
367	5	2	0	0	0	9.49	103	100		98	102	92	94	87	104	112		114
369	5	2	0	0	0	9.57	107	114			112	115			101	111		
376	5	2	0	0	0	9.86	94	79		91	98	82	87	100	90	75		82
377	5	2	0	0	0	10.45	112			103	102		122	98	124			109
398	5	2	0	1	0	9.49	95	87			74	73	77		110	94		
373	5	2	0	0	1	10.03	102	90		115	104	93	103	117	99	85		113
392	5	2	0	0	1	10.03	120	118		128	147	150	149	153	95	91		109
374	5	2	0	0	1	9.49	91	108			91	95	102		91	128		
350	5	2	0	0	1	9.53	118	104		106	132	127	128	128	104	84		91
378	5	2	0	0	1	9.53	87	101		95	98	105	104	113	73	97		79

ID#	G	A	T	M	S	Age	Total IQ				Verbal IQ				Reasoning IQ			
							1	2	3	4	1	2	3	4	1	2	3	4
379	5	2	0	0	1	9.61	104	97		101	103	114	104	109	106	84		93
383	5	2	0	0	1	9.78	112	124		148	120	129	121	140	99	109		156
365	5	2	0	0	1	9.61	140	163		150	146	160		142	135	168		158
386	5	2	0	0	1	9.90	92	85		88	100	98	95	90	77	66		86
385	5	2	1	1	0	10.24	79	90		94	84	87	91	100	71	87		85
368	5	2	1	0	1	9.95	99	127			108	122			91	132		
364	5	2	1	0	1	9.95	85	97		124	105	101	104	138	63	93		109
390	5	2	1	0	1	10.78	61	79		81	65	69	60	75	54	83		90
354	5	3	0	0	0	9.78	117	127		127	120	129		140	112	124		115
341	5	3	0	0	0	10.03	93	90		94	97	97		105	89	80		83
355	5	3	0	0	0	9.45	98	94	92	111	94	103	105	114	105	83	78	107
334	5	3	0	0	0	10.24	119		123	122	115		116	131	123		131	115
361	5	3	0	0	0	10.24	101	103	123	113	95	87	101	97	107	128	181	142
348	5	3	0	0	0	9.57	96	107	119	109	85	107	111	106	109	107	132	113
333	5	3	0	0	0	10.24	98	114	118	115	82	98	105	103	123	140	136	131
335	5	3	0	0	0	10.24	125	115	131	103	156	140	146	100	104	98	116	107
344	5	3	0	0	0	10.11	106	104	113		103	109	113		109	99	113	
345	5	3	0	0	0	10.45	94	115	118		82	106	114		109	126	122	
351	5	3	0	0	0	9.70	126	142	171	157	165	197	249	229	102	114	150	137
357	5	3	0	0	0	10.61	106	112	141		115	115	158		96	108	127	
359	5	3	0	0	0	10.32	120	136	148	125	101	118	153	125	155	167	141	125
337	5	3	0	0	0	10.15	133	145	143	132	158	152	147	136	116	136	137	127
340	5	3	0	0	1	9.95	101	110	111	118	115	115	123	118	89	104	100	118
349	5	3	0	0	1	9.65	99	111	113	110	108	136	138	132	89	94	96	95
353	5	3	0	0	1	10.24	90	104	123	131	95	124	125	135	84	89	120	126
331	5	3	0	0	1	9.32	118	168			115	160			122	183		
347	5	3	0	0	1	10.32	106			103	136		130	120	83			90
360	5	3	0	1	1	10.65	107	104	135	133	100	119	141	130	115	92	126	137
336	5	3	1	0	0	10.28	123	146	145	158	156	150	153	218	101	140	136	142
358	5	3	1	0	0	10.20	110	108	116		116	101	109		102	116	125	
342	5	3	1	0	0	9.57	139	123			154	127			127	119		

ID#	G	A	T	M	S	Age	Total IQ 1	2	3	4	Verbal IQ 1	2	3	4	Reasoning IQ 1	2	3	4
352	5	3	1	0	1	9.32	139	157	159	158	127	160	177	163	158	153	142	154
356	5	3	1	0	1	9.95	110	93	100	90	111	96	107	98	108	91	93	83
338	5	3	1	1	1	9.82	111	140	165	147	116	133	160	156	106	146	169	136
470	6	1	0	0	0	12.28	62	58	63		59	57	59		66	59	67	
471	6	1	0	0	0	10.86	79	76	81		87	79	88		70	73	73	
474	6	1	0	0	0	11.90	87		93		76		83		103		109	
469	6	1	0	0	1	11.11	81	91	91		87	93	104		73	88	80	
460	6	1	0	0	1	12.78	83	79	80		92	82	94		73	76	67	
446	6	1	0	0	1	12.32	68	75	71		64	70	70		72	82	73	
466	6	1	0	0	1	11.86	112	114			138	138			93	97		
468	6	1	0	0	1	11.74	60	59			69	61			52	55		
458	6	1	0	0	1	12.03	80	96	100		86	102	97		71	90	104	
441	6	1	0	0	1	11.07	107	97	106		126	107	122		92	87	94	
455	6	1	0	0	1	10.99	106	110	123		111	111	123		100	108	123	
434	6	1	0	1	1	11.11	96	87	104		99	93	104		94	82	104	
444	6	1	0	1	1	12.99	68	66	76		61	53	64		75	75	100	
473	6	1	0	1	1	11.32	76	79			72	66			80	95		
463	6	1	0	1	1	11.42	74	84	85		69	84	86		78	84	84	
465	6	1	1	1	0	11.99	75	81	82		58	72	69		95	96	108	
459	6	1	1	0	1	11.82	90	91	94		107	118	125		75	73	76	
461	6	1	1	1	1	11.03	93	91	101		114	85	108		76	97	95	
472	6	2	0	0	0	10.53	93	91	102		80	83	84		112	102	139	
445	6	2	0	0	0	10.90	112	111	120		112	109	134		112	112	109	
439	6	2	0	0	0	11.32	104	102	127		104	113	124		104	92	130	
423	6	2	0	0	0	11.42	100	106	118		89	98	105		114	116	139	
443	6	2	0	0	0	11.15	92	92	92		93	96	84		91	88	104	
449	6	2	0	0	0	10.86	94	86	86		99	86	88		89	86	83	
440	6	2	0	0	0	10.95	110	105			108	98			111	116		
438	6	2	0	1	0	11.07	84	78	86		80	81	86		89	76	86	
447	6	2	0	0	1	10.99	100	137			118	141			85	131		
457	6	2	0	0	1	10.61	98	101	112		93	84	90		104	130	176	
442	6	2	0	0	1	12.32	108	123	110		106	118	101		110	133	120	
433	6	2	0	0	1	10.32	107	96	119		131	127	180		88	74	92	

ID#	G	A	T	M	S	Age	Total IQ				Verbal IQ				Reasoning IQ			
							1	2	3	4	1	2	3	4	1	2	3	4
437	6	2	0	0	1	10.86	122	120	129		141	127	135		109	113	124	
450	6	2	0	0	1	10.99	88	93	96		90	89	89		85	98	105	
464	6	2	0	1	1	12.82	77	85	90		74	87	85		81	85	98	
467	6	2	1	0	0	11.28	75	86	85		59	70	72		95	113	106	
451	6	2	1	1	0	11.36	86	76	81		84	76	75		90	77	89	
454	6	2	1	0	1	11.28	88	99	103		86	102	96		90	95	110	
435	6	2	1	0	1	11.24	140	149	158		146	171	217		131	134	141	
412	6	3	0	0	0	9.70	152	148	150		144	125	153		158	197	143	
402	6	3	0	0	0	11.03	125	120	148		127	126	144		122	115	152	
425	6	3	0	0	0	11.07	108	122	139		103	104	127		117	156	169	
429	6	3	0	0	0	10.49	104	128	133		153	155	178		75	109	110	
420	6	3	0	0	0	10.99	111	116	133		115	137	133		107	101	133	
430	6	3	0	0	0	10.65	107	119	120		98	108	101		118	135	157	
405	6	3	0	0	0	11.07	106	109	123		106	111	127		107	107	122	
427	6	3	0	0	0	11.11	124	119	123		113	130	121		138	110	126	
415	6	3	0	0	0	10.61	98		110		104		127		93		98	
417	6	3	0	0	0	10.95	93	99	109		89	101	99		98	98	123	
418	6	3	0	0	0	11.36	103	122	109		107	127	99		97	116	124	
419	6	3	0	0	0	11.42	94	98	109		94	97	94		94	101	139	
431	6	3	0	0	0	10.90	98	99	109		108	101	118		89	95	103	
424	6	3	0	0	0	11.28	115	115	114		115	113	110		115	117	120	
428	6	3	0	0	0	11.28	81	89			79	90			82	87		
408	6	3	0	0	0	10.99	82	85	86		83	87	85		81	83	87	
422	6	3	0	1	0	10.95	98	123	141		98	132	134		98	116	153	
426	6	3	0	0	1	10.70	139	152	148		162	161	156		121	141	137	
416	6	3	0	0	1	11.24	95	109	103		93	102	96		98	118	110	
456	6	3	0	1	1	11.65	80	91	105		89	106	121		70	79	93	
409	6	3	0	1	1	11.15	115	139	126		121	139	135		109	135	115	
411	6	3	1	0	0	11.45	107	123	123		110	121	132		103	126	112	
413	6	3	1	0	0	10.74	104	112	119		92	103	100		121	123	174	
410	6	3	1	0	0	10.82	128	130	135		129	133	135		125	128	135	
421	6	3	1	0	1	11.03	125	134	140		145	148	144		111	120	133	
404	6	3	1	0	1	11.03	118	127			118	131			118	126		
406	6	3	1	0	1	10.49	95	108			120	116			75	99		