DOCUMENT RESUME

ED 213 788 UD 022 127

AUTHOR Hoepfner, Ralph, Ed.
TITLE Substudies on Allocation and Targeting of Funds and Services, Assessment of Student Growth, and Effects of Attrition. Technical Report #13 from the Study of the Sustaining Effects of Compensatory Education on Basic Skills.
INSTITUTION System Development Corp., Santa Monica, Calif.
SPONS AGENCY Office of Program Evaluation (ED), Washington, D.C.
PUB DATE May 81
CONTRACT 300-75-0332
NOTE 319p.; Some tables marginally legible. For related documents see ED 146 182-183, ED 155 300, ED 163 128, and UD 022 122-128.

EDRS PRICE MF01/PC13 Plus Postage.
DESCRIPTORS Academic Achievement; *Achievement Rating; Admission Criteria; *Affective Measures; *Compensatory Education; Elementary Education; Elementary School Students; Evaluation Methods; *Financial Support; Low Income Groups; Minority Groups; *Poverty Areas; Program Effectiveness; School Demography; *Student Attrition
IDENTIFIERS *Elementary Secondary Education Act Title I

ABSTRACT

This study of Compensatory Education is divided into four parts: allocation of Title I funds and services, achievement and Compensatory Education, measurement of student growth, and the effects of attrition. Part one examines poverty concentrations to determine which schools and districts are eligible for Title I or for concentration grants. The methods for determining a school's poverty level are also presented. Part two discusses how and how well Title I and Compensatory Education are distributed at the student level. The results of student achievement test scores are compared with teachers' evaluation of students' needs for Compensatory Education. The procedures used in the selection and development of measures of academic growth, achievement, and affective growth are described in part three. Part four describes the reduction in the sample of schools in the second year of the study after funding cutbacks. Also reported is the incidence of student attrition compared over several characteristics including achievement level and minority status. (Author/JCD)

REPORT 13:
SUBSTUDIES ON ALLOCATION AND TARGETING OF FUNDS AND SERVICES, ASSESSMENT OF STUDENT GROWTH, AND EFFECTS OF ATTRITION

Ralph Hoepfner, Editor

Technical Report No. 13 From the Study of the Sustaining Effects of Compensatory Education on Basic Skills

Prepared for the
Office of Program Evaluation
U.S. Department of Education

by

System Development Corporation
2500 Colorado Avenue
Santa Monica, California 90406

May 1981

REPORTS IN THE STUDY OF SUSTAINING EFFECTS OF COMPENSATORY EDUCATION SERIES

1. Hoepfner, Ralph; Zagorski, Henry; and Wellisch, Jean. The Sample for the Sustaining Effects Study and Projections of its Characteristics to the National Population. System Development Corporation, Santa Monica, California, June 1977.

2. Breglio, Vincent J.; Hinckley, Ronald H.; and Beal, Richard S. Students' Economic and Educational Status and Selection for Compensatory Education. Decima Research, Santa Ana, California, January 1978.

3. Hinckley, Ronald H.; Beal, Richard S.; and Breglio, Vincent J. Student Economic and Educational Status and Receipt of Educational Services. Decima Research, Santa Ana, California, June 1978.

4. Hinckley, Ronald H. (Editor); Beal, Richard S.; Breglio, Vincent J.; Haertel, Edward H., and Wiley, David E. Student Home Environment, Educational Achievement and Compensatory Education. Decima Research, Santa Ana, California, January 1979.

5. Wang, Ming-mei; Hoepfner, Ralph; Zagorski, Henry; Hemenway, Judith A.; Brown, Deborah S., and Bear, Moraye B. The Nature and Recipients of Compensatory Education. System Development Corporation, Santa Monica, California, September 1978.

6. Haggart, Sue A.; Klibanoff, Leonard S.; Sumner, Gerald C.; and Williams, Richards S. Resource Analysis of Compensatory Education. RMC Research Corporation, Santa Monica, California, October 1978.

7. Sumner, Gerald C.; Klibanoff, Leonard S.; and Haggart, Sue A. An Analysis of the Cost and Effectiveness of Compensatory Education. RMC Research Corporation, Santa Monica, California, August 1979.

8. Klibanoff, Leonard S., and Haggart, Sue A. Summer Growth and the Effectiveness of Summer School. RMC Research Corporation, Santa Monica, California, June 1980.

9. Hemenway, Judith A.; Wang, Ming-mei; Kenoyer, Charles E.; Hoepfner, Ralph; Bear, Moraye B.; and Smith, Gerrie. The Measures and Variables in the Sustaining Effects Study. System Development Corporation, Santa Monica, California, December 1978.

9A. The SES Staff. A Compilation of the Instruments Used in the Sustaining Effects Study. System Development Corporation, Santa Monica, California, April 1979.

11. Kenoyer, Charles E.; Cooper, Deborah M.; Saxton, D. Eugene; and Hoepfner, Ralph.

The Effects of Discontinuing Compensatory-Education Services. System Development Corporation, Santa Monica, California, February 1981.

13. Hoepfner, Ralph (Editor). Substudies on Allocation and Targeting of Funds and Services, Assessment of Student Growth, and Effects of Attrition. System Development Corporation, Santa Monica, California, May 1981.

The work reported in this series was performed under Contract No. OE 300-75-0332 with the U.S. Department of Education. The Department of Education encourages contractors to express their professional judgments in reports. The opinions, conclusions, and recommendations in these reports are those of the authors and do not necessarily represent Office of Education positions or policy.

THE STAFF OF THE STUDY OF SUSTAINING EFFECTS

SYSTEM DEVELOPMENT CORPORATION:

Project Management. Launor F. Carter, Director (1975–1980); Ralph Hoepfner, Director (1981–); Raymond B. Stewart, Associate Director (1975–1979); Thomas F. Collins, Assistant Director for Planning and Control (1975–1978); Robert L. Lamoureux (1978); Marilyn Stevenson (1975–1977); Patrice Metro (1977–1979).

Design and Analysis. Ralph Hoepfner, Manager (1975–); Henry J. Zagorski (1975–); Ming-mei Wang (1976–); Charles E. Kenoyer (1975–1976, 1978–1979); Joel Moskowitz (1977–1978); Ray W. Baker (1976–1978); Mary Jean Austin (1976–1980); John McCall (1977); Lewis Zamora (1976–1977); Deborah S. Brown (1977–1978); Judith A. Hemenway (1977–1979); Thomas E. Locke (1977–1978); Wallace S. Tai (1977–1978); Marino Giustino (1978); Moraye B. Bear (1977–1980); Lawrence A. Jordan (1978–1979); Deborah Cooper (1978–1981); Edward J. Colon (1978–1979); Gerrie Smith (1978–1979); Eugene Won (1979–); Jon Conklin (1979–).

Instrumentation and School Characteristics. Jean B. Wellisch, Manager (1975–1977); Ralph Melaragno (1975–1976); Ronald A. Carriere (1975–1977).

Field Operations. Dean R. Lee, Manager (1975–); Cleone L. Geddes (1975–1980); Thomas A. Ramirez (1975–1979); Cynthia Hull (1976–1977); John R. Shiban (1975–1976).

Successful Sites Study. Dean R. Lee, Manager (1977–); Gary A. Duck (1977–1978); Lee H. Poynor (1977–1981); Ronald A. Carriere (1978–1981); Anne H. MacQueen (1978–); Fred E. Cromer (1978); Miles S. Rogers (1979–).

Distribution and Scoring. Patrick F. Carnes, Manager (1976–1979); Carl Molitor (1976–1977); Loren Southworth (1976–1977); Stanley A. Dounn (1976–1979); Herb Smith (1977–1978).

Editorial and Publications. Pamela Doherty, Manager (1978–); Bruce Severy (1980–); Frank Tierney (1978–1981); Kean Mantius (1975–1977); Una Vere Katter (1976–1978).

DECIMA RESEARCH:

Participation Study. Vincent J. Breglio, Manager (1975–1979); Ronald H. Hinckley (1975–1979); Richard S. Beal (1977–1978); Timothy McGrew (1975–1977); Claudia Washburn (1976–1977); S. Kay Lavish (1975–1977); Barbara Openshaw (1975–1977).

RMC RESEARCH CORPORATION:

Cost/Effectiveness and Summer Study. Sue A. Haggart, Manager (1975 – 1979); Leonard S. Klibanoff (1975 – 1979); Nabeel Al-Salam (1975 – 1977); Laurence A. Dougharty (1975 – 1977); Gerald C. Sumner (1977 – 1979); Richards S. Williams (1977 – 1979); Rita Steinmetz (1977 – 1979).

PACIFIC CONSULTANTS:

Functional Effectiveness Measures. Mary Spencer, Manager (1975 – 1976); David Bessemer (1975 – 1976); Nicolas Fedan (1975 – 1976); Bobby Offutt (1975 – 1976).

U.S. OFFICE OF EDUCATION:

Office of Planning, Budgeting and Evaluation. Janice K. Anderson, Project Officer (1979 –) and Deputy Project Officer (1975 – 1979); George W. Mayeske, Project Officer (1975 – 1979); Kathryn E. Crossley, Administrative Assistant (1975 –).

Division of Education for the Disadvantaged. Paul Miller (1975 –); William Lobosco (1975 –); Velma James (1975 –).

SUSTAINING EFFECTS STUDY ADVISORY GROUPS AND PANELS

Policy Advisory Group. Edward Boddy (1975–); Oteal Bowen (1975–); Alan Davitt (1977–1979); Leo Doherty (1975–); Shirley Foster (1975–); Joe Hansen (1975–1976); Walter Hathaway (1975–1976); Rio Rita Jackson (1975–); Robert Lamborn (1977); Celia Martin (1975–); Roberta Martinez (1975–); Lucy Matos (1975–); Miles Myers (1975–); Thomas Ralls (1975–); Stanley Rumbaugh (1975–); Alexander Sergienko (1975–); Kenneth Smith (1977); Patricia Williams (1975–).

Research Advisory Group. Donald Campbell (1975–); William Cooley (1975–); Rex Fortune (1975–); William Michael (1975–); David Wiley (1975–); Alex Mood (1975–1980); Melvin Novick (1975–1978).

Panel on Achievement Tests. William Coffman (1975–1976), Melvin Seeman (1975–1976); Floraline Stevens (1975–1976); Ralph Tyler (1975–1976); James Vasquez (1975–1976).

Panel on Bias in Evaluation. Thomas F. Pettigrew (1975–1976); Benjamin J. McCullough (1975–1976); Daniel O'Neal (1975–1976); Frank Navarette (1975–1976); Ricardo Chapa (1975–1976); Florida Catchins Hyde (1975–1976); Dorothy Peterson (1975–1976); Takako Okubo (1975–1976); Elma Bertram (1975–1976); Don Manalili (1975–1976).

Panel on Functional Literacy. Marilyn Litchman (1975–1976); James Saunders (1975–1976); Thomas Sticht (1975–1976); James Vasquez (1975–1976).

Panel on Affective Measures. Joyce Levy Epstein (1975–1976); John Kitsuse (1975–1976); Melvin Seeman (1975–1976); James Vasquez (1975–1976).

Panel on Innovative Compensatory Projects. Mary Bentzen (1975–1976); William Coffman (1975–1976); William Georgiades (1975–1976); John Peper (1975–1976); Ralph Tyler (1975–1976).

TABLE OF CONTENTS

TABLE OF CONTENTS (Continued)

TABLE OF CONTENTS (Continued)

PART IV. THE EFFECTS OF ATTRITION

LIST OF TABLES

LIST OF TABLES (Continued)

LIST OF TABLES (Continued)

LIST OF TABLES (Continued)

LIST OF TABLES (Continued)

LIST OF FIGURES

A GENERAL INTRODUCTION TO THE SUSTAINING EFFECTS STUDY AND AN OVERVIEW OF THE PRESENT REPORT

GENERAL INTRODUCTION

In response to questions about education policies, SDC is studying compensatory education (CE): its nature, quantity, and environment; its sustained effects; and its generality, in a large study called: The Sustaining Effects Study. This thorough study will result in a series of reports from the following substudies:

The Longitudinal Study. In the Longitudinal Study, the growth of children in reading, math, functional literacy, and attitudes toward school were assessed in the fall and spring for three consecutive years. The amount and kind of instruction in reading and math were also determined for each student, and, in addition, teachers and principals reported on their philosophies and practices of instruction. Thus, it was possible not only to assess student growth over a three-year period, but to relate that growth to certain specifics of instruction.

The schools in the study were drawn from three different groups. The REPRESENTATIVE SAMPLE of schools is a sample carefully drawn to represent all of the nation's public schools that have some of the grades one through six. A second group of schools, the COMPARISON SAMPLE, is composed of schools that have large proportions of students from poor homes but do not receive special funds to offer CE services. The third group is the NOMINATED SAMPLE, composed of schools nominated because their educational programs had promise of being effective for low-achieving students. During the first year of the study, data were collected from 328 schools and about 118,000 students.

The Cost/Effectiveness Study. Information was obtained on the resources and services to which each student was exposed during reading and math instruction. Cost estimates were generated on the basis of this information. Because the effectiveness of the instructional programs is to be determined as a product of the Longitudinal Study, it will be possible to relate the effectiveness of programs to their costs.

The Participation Study. The purpose of the Participation Study was to determine the relationships among economic status, educational need, and instructional services received. The educational achievement of the students and the services they received were obtained in the Longitudinal Study, and the refined measures of economic status were obtained in the Participation Study. Visits were made to the homes of over 15,000 randomly selected students from the schools in the first-year REPRESENTATIVE SAMPLE. During the visits, information was collected on the economic level of the home and on the parents' attitudes toward their children's school and learning experiences. Thus, the level of student achievement and services could be related to the economic level of a student's home.

The Summer Study. The Sustaining Effects Study also examined the effectiveness and cost-effectiveness of summer-school programs. Information about the summer school experiences of students was combined with other data. A 'resource-cost' model, developed for the regular-year, cost-effectiveness study, was adapted to the needs of the summer-school study.

Successful Practices in High-Poverty Schools. This study is intended to identify and describe instructional practices and contexts that appear to be effective in raising the reading and math achievements of educationally disadvantaged students. In-depth observational and interview data were collected from 55 schools that are participating in the study.

THE REPORT SERIES

The major findings of the reports already published are discussed briefly below, along with references to the specific reports from the study that have addressed them.

A Description of the Samples for the Sustaining Effects Study and the Nation's Elementary Schools. In order to understand the findings of this study, it is essential to become familiar with the characteristics of the samples used and their capabilities of providing generalizations to the population of the nation's schools. Technical Report 1 (Hoepfner, Zagorski, and Wellisch, 1977) describes in detail the samples and how they were formed. It also presents the results of a survey of 4,750 public schools with grades in the 1-6 range by projecting the data to the nation. These projections accurately describe the nation's elementary schools in terms of the characteristics of the schools, the kinds of services the schools provide to students, and the characteristics of the students. The interrelationships among these characteristics are also analyzed.

The different kinds of samples have been explained earlier in this review. Some results concerning the characteristics of the nation's public schools are summarized below:

- *Enrollment, Urbanism, and Achievement.* The total grade 1-6 enrollment in the 1975-76 school year was estimated at about 21 million students. There is a moderately strong relationship between school enrollment and degree of urbanism, with large cities having larger schools than rural areas. Student achievement is related to urbanism in a complex way; in general, there are proportionally more large city than rural schools that have more than half of their students with achievement levels at least one year below grade.

- *Compensatory-Education Funds, School Characteristics, and Achievement.* About two-thirds of the nation's elementary schools received Title I funds, and about one-fifth received no compensatory funds from any sources. There is little relationship between the receipt of compensatory funds and the size of a school. However, small-city and rural schools tend to receive such funds more frequently than do large-city schools. As expected, schools with high concentrations of poor students tend to receive compensatory funds more often than do schools with low concentrations.

Similarly, schools with higher percentages of low-achieving students are more likely to receive compensatory funds.

- *Achievement and the Concentration of Poor and Minority Students.* There is a strong association between percentages of low-achieving students and concentrations of poor and minority students.

- *School's Grade Span.* Generally, the grade span in the school shows weak relationships with the size of school, the degree of urbanism, and the concentrations of low-achieving, poor, and minority students.

- *Stability of Student Body.* Schools tend to have less stability in their student bodies as the size of the school increases, and there tends to be less stability in large cities. Similarly, stability decreases as concentrations of poor, minority, and low-achieving students increase.

- *Availability of Summer Schools.* Fifty-one percent of the nation's schools with grades 1-6 have summer-school programs available for their students. Larger schools provide summer-school programs more frequently than smaller schools. There is practically no relationship between the availability of summer school and a school's level of poverty, minority concentration, or student achievement level.

A Description of Student Selection for Compensatory Services As It Relates to Economic Status and Academic Achievement. The Education Amendments of 1974 require several studies to inform Congress concerning who does and who does not receive Title I services and how selection for such services is related to the economic status of the family and the academic performance of the child (Section 417 of the General Education Provision Act). In addition, the federal program administrators want to know the differences between the services received by economically and educationally deprived children and those by non-deprived children, and the relationship between academic achievement and the home environment.

These questions were addressed in Technical Reports 2 (Breglio, Hinckley, and Beal, 1978), 3 (Hinckley, Beal, and Breglio, 1978), and 4 (Hinckley, Ed.; Beal, Breglio, Haertel, and Wiley, 1979). A brief summary of answers to the questions is provided below:

- About 29 percent of poor students participate in Title I compared to about 11 percent of the non-poor (Report 2). Looking at CE in general, about 40 percent of the poor students and about 21 percent of the non-poor participate. From these findings, we can see that proportionally more poor students participate in the services than non-poor.

- Using a grade-equivalent metric (one year below expectation for the student's current grade) as the definition for educational disadvantage, about 31 percent of low-achieving students participate in Title I, while only 10 percent of regular-achieving students do (Report 2). For CE in general, the percentages are 46 for low-achievers and

19 for regular-achievers. Among the regular-achievers who participate in CE, many score below the national median on achievement tests.

- Participation rates for Title I and for CE in general, are the highest for students who are both economically and educationally disadvantaged (Report 2). Forty-one percent of those students participate in Title I, and 54 percent participate in CE in general. Participation rates are next highest for students who are educationally but not economically needy (26 and 41 percent, respectively), and next highest for students economically but not educationally needy (20 and 28 percent, respectively). Only 7 percent of the students who are neither educationally nor economically needy participate in Title I (15 percent for CE in general). These participation rates were interpreted as indicating that the then-current allocation procedures were being complied with, and the intentions of the law were being met fairly.

- In comparison to non-poor students, poor students receive more hours of instruction per year with *special* teachers, more hours of instruction in medium- and small-sized groups, fewer hours of independent study, and more non-academic services, such as guidance, counseling, health and nutrition (Report 3). The differences are even stronger when poor Title I students are compared to others. Therefore, we can conclude that the distribution of educational services is in line with the intention of the laws and regulations.

- Two aspects of the children's home environments bore significant and consistent relations to achievement. the amount of reading done at home and the educational attainment of the head of the household. Other variables, such as family size, TV-watching behavior, and type of living quarters were not consistently related to student achievement (Report 4). Although most parents (67 percent) know whether their children's schools have special programs for low-achieving students, few (40 percent) know of Title I and even fewer know of or participate in local governance of the Title I programs. Poor parents, in general, are less involved in their children's educational programs, have lower expectations of their children's attainments, and give lower ratings to the quality of their children's educations, but still perceive Title I and other CE programs as being helpful.

Description of the Nature of CE Programs, Characteristics of Participating Students, Schools, and Educational Services. The Participation Study deals almost exclusively with what has been called 'selection for CE or Title I services,' without examining too closely what such programs really are and how they differ from the programs regularly offered by the schools. Before we could draw any conclusions about participation in a CE program and the educational progress of students, we had to be assured that there really was a program that was distinct, could be specified in some way, and had a reasonable chance of making an impact. As will be seen, not only did we analyze data on the basis of program participation, but we also considered the actual services received in order to address directly the possible differences between intentions and actualities.

Based on the analyses of data obtained from about 81,500 students in the Representative Sample of schools, Technical Report 5 (Wang, Hoepfner, Zagorski, Hemenway, Brown, and Bear, 1978) provides the following important conclusions:

- Students participating in CE are lower achievers (mean score at the 32nd percentile) than non-participants (53rd percentile). Seventy percent of the participants were judged by their teachers as needing CE, while only 19 percent of those not participating were so judged. More minority students participate in CE, proportionately, than white students, but participation in CE has little relationship with student attitudes towards school, early school experiences, summer experiences, or the involvement of their parents in their educational programs.

- Minority, poor, and low-achieving students tend to receive more hours of instruction in smaller groups and by special teachers, and to receive more non-academic services, but their attendance rates are generally lower, too, so that they do not take maximum advantage of the special services provided.

- The useful predictors of whether or not a student is selected to receive CE are his/her teacher's judgment of need and participation in CE in the previous year. When these variables are considered, achievement scores, non-English language spoken in the home, and economic status contribute little more to the prediction.

- About two-thirds of the students participating in CE in 1975-76 participated in the 1976-77 school year also.

- CE students in general and Title I students in particular receive more total hours of instruction per year than non-CE students. The CE students also receive more hours of instruction from special teachers. Among CE students, Title I students receive the greatest number of hours of instruction, more frequently with special teachers, and in small instructional groups. There are no significant and consistent differences between CE students and non-CE students with regard to their teachers' instructional subgrouping practices, the use of lesson plans, the extent of individualization of instruction, the frequency of feedback, or the assignment of homework.

- Students receive 5 to 9 hours of reading instruction per week, decreasing steadily with higher grades, and between 5 to 6 hours of math instruction per week, fairly constant over all grades.

- CE services are delivered during regular instructional hours with different kinds of activities for the participants (so that, in effect, they 'miss' some regular instruction received by non-participating peers).

- Title I schools have higher average per-participant CE expenditures in reading and math than do schools with other CE programs. The average Title I per-participant expenditure is about 35 percent of the average per-pupil regular (base) expenditure.

- Schools providing CE generally have higher concentrations of poor and low-achieving students, and students with less educated parents. These schools have greater administrative and instructional control by their districts and have higher staff-to-student ratios.

- Schools that select higher percentages of regular-achieving students for CE services have larger percentages of minority and poor students, probably reflecting their tendency for the saturation of CE programs.

- Most districts use counts of students receiving reduced-price lunches and counts of families receiving AFDC to determine school eligibility for compensatory funds, while most schools select students on the basis of standardized achievement tests, frequently augmented by teacher judgments. Similar selection criteria are employed by non-public schools.

Cost-Effectiveness of Compensatory Education. In its deliberations for the reauthorization of Title I, and in annual appropriation hearings, members of Congress also wanted information on the effectiveness of the Title I program relative to its cost. While it appears eminently sensible to ask the question of cost-effectiveness, it is difficult to provide the answers in a manner that will be interpreted correctly.

In the study of the cost-effectiveness of CE, efforts were made to preclude enigmatic conclusions and, at the same time, to make cost estimates on a sounder basis than in the past. In Technical Report 6, Haggart, Klibanoff, Sumner, and Williams (1978) develop and present a resource-cost model that translates educational resources for each student into estimates of average or standard dollar costs for his/her instructional program. The overall strategy for estimating costs is to provide an index that represents the labor-intensity of services without being confounded with regional price differentials, different accounting methods, etc.

Using the resource costs, CE students in general, and Title I students in particular, were found to be offered substantially higher levels of educational resources, and hence more costly programs. Participation in CE differentiates the resource costs for services offered much more than do poverty, achievement level, race, or any other characteristics.

In Technical Report 7, Sumner, Klibanoff, and Haggart (1979) related resource costs to achievement to arrive at an index of cost-effectiveness. Because of the low-achievement levels of the children participating in CE, and their relatively slow rates of achievement growth, the increased costs associated with CE appeared to be misspent (in the same way that money for severly ill and terminal patients appears to be not as effectively spent as it is for mildly ill patients). It is important to point out, however, that the appearance may not tell the true story. Because we cannot obtain truly appropriate comparison groups, we do not know what would have happened to the achievement growth of the CE students if they had not participated. Based on the comparison groups we could form, however, CE programs did not appear to have an advantage over regular programs in terms of cost-effectiveness.

The Effectiveness of Summer-School Programs. The study has also examined the results of attendance at summer school, because members of Congress and program administrators wanted to know if such attendance helps prevent the presumed progressive academic-deficit of low-achieving students. If attendance at summer school were to have a positive academic effect (insofar as that the attendees do not 'fall back' to their achievement levels of previous years), then summer programs could be considered as a means of sustaining school-year growth.

Technical Report 8 (Klibanoff and Haggart, 1980) shows that attendance at summer school has little or no effect on academic growth, especially on low-achieving students. Because the findings are based on the study of summer schools as they presently exist (and the evidence is strong that they do not offer intensive academic experiences), the non-positive findings should not be interpreted as an indictment of summer school, as such, but an evaluation of the way they are presently organized and funded. Nevertheless, when instructional services delivered in summer schools were investigated, none seemed particularly effective in improving achievement growth.

In the same report, the authors also addressed the hypothesis of 'summer drop-off,' a hypothesis advanced to explain the presumed widening achievement gap between regular and CE students. Essentially, this hypothesis states that CE students lose much more of their previous year's learning during the summer recess than do regular students. Data collected in the study fail to support the summer drop-off hypothesis. CE students do not suffer an absolute 'drop-off' (although their achievement growth over the summer is less than that for regular students, as in the school year). In any event, attendance at summer school does not have much of an effect.

Technical Report 9 is a resource book. It identifies all the variables and composites that have been selected or devised for use in the Sustaining Effects Study. All measures and scales are described and rationalized. In addition, Report 9A serves as a companion volume that contains copies of all of the data-collection instruments in the study, except for a few that are constrained by copyright.

Compensatory Services and Educational Development in the School Year. Technical Report 10 addresses the effects of compensatory services on student development during the school period. It also examines the instructional services and major dimensions of the educational process to describe the characteristics of programs that are effective in raising achievement levels. The analysis is based on the first-year data of the study. Similar investigations will continue in subsequent reports.

Studies Still to be Done. The remaining reports, yet to come from the study, will address the general effects of educational practices on raising achievement levels, with special attention paid to the practices found in CE programs in general and in Title I programs in particular. Impact analyses will either be based on three-year longitudinal data or on in-depth observations and interviews. The extensive achievement data collected from overlapping cohorts of students in the three years will be utilized to describe the patterns of educational growth over the years for various groups of CE and non-CE students. Analyses of the three-year longitudinal data will allow us to examine the sustained effects of CE and help us determine if the presumed phenomenon of gap-widening between the disadvantaged and non-disadvantaged students indeed exists.

OVERVIEW

This report includes a collection of substudies that are part of the Sustaining Effects Study (SES). The substudies apply selected data to specific policy issues or investigate in greater depth certain aspects of our data that were not fully understood. The findings, we believe, are of general interest and are of sufficient value to warrant publication. The substudies are organized, as chapters, into four topics: allocation of Title I funds and services (Part I), achievement and compensatory education (Part II), measuring student growth (Part III), and the effects of attrition (Part IV).

ALLOCATION OF TITLE I FUNDS AND SERVICES

Part I contains five chapters of small studies and analyses especially requested by the Department of Education to provide useful information to the participants in the then current Congressional hearings on the reauthorization of Title I in 1978. The first chapter addresses poverty concentrations in districts and schools in order to provide estimates of the potential incidence of 'target areas' and 'concentration grants.' All schools in 'target areas' of relatively high poverty would become automatically eligible to participate in Title I. Districts with very high poverty concentrations are eligible for 'concentration grants.' Below, we make projections of the national distribution of poor districts and schools over regions and categories of urbanism. The distributions of schools with poverty concentrations at or above specified criteria for eligibility are also presented.

Percentages of High-Poverty Schools by Region and Urbanism

Region / Urbanism	Potential Criteria for Target Areas: 10% + Poverty*	Potential Criteria for Target Areas: 20% + Poverty	Potential Criteria for Target Areas: 30% + Poverty	Potential Criteria for Concentration Grants: 70% + Poverty	Potential Criteria for Concentration Grants: 80% + Poverty
New England	7.3	6.7	6.3	7.4	5.3
Metropolitan Northeast	7.1	6.7	6.6	8.0	9.1
Mid-Atlantic	10.4	10.6	10.9	9.0	6.4
Southeast	18.4	22.9	27.9	28.3	28.4
North Midwest	18.3	15.0	12.6	10.3	12.1
South Central	13.1	15.6	17.4	20.5	21.6
Central Midwest	7.3	5.8	4.3	3.1	2.9
North Central	4.8	3.8	2.8	2.2	2.8
Pacific Southwest	9.2	9.3	8.7	9.2	8.7
Pacific Northwest	4.2	3.6	2.4	2.1	2.7
City of over 500,000	8.3	10.6	13.5	29.2	35.2
City of 200,000-500,000	4.3	4.9	5.4	7.4	7.6
City of 50,000-200,000	10.8	11.1	12.3	11.1	10.1
Suburb of city	9.1	6.4	4.7	1.5	1.2
City under 50,000	28.2	27.3	25.5	17.8	14.0
Rural area near city	10.5	8.8	7.3	3.6	2.1
Rural area not near city	28.8	30.8	31.3	29.3	29.8

*The criteria are based on principals' estimates of the percentages of students enrolled in the school who meet Title I poverty criteria or are eligible to receive free lunches.

Because of their poverty, the Southeast and South Central regions, together with large cities and remote rural areas, would gain in the number of schools automatically eligible, as the required poverty rate increases. On the other hand, if the required poverty rate is lower, the North Midwest and suburbs, small cities, and rural areas near cities would automatically gain eligible schools. Higher poverty rates for concentration grants increase the numbers of schools eligible in the Metropolitan Northeast and the North Midwest regions and in large cities, but decrease them in the New England and Mid-Atlantic regions and in small cities. (Partially based on these data, the criteria specified in the 1978 reauthorization were 20 percent for target areas and 75 percent for concentration grants.)

In the second chapter we study the results of different methods for determining a school's poverty level in order to check the validity of data in the SES, and also to pinpoint any weakness in methodology. Five estimates of the concentration of students from poor families are compared. The estimates, made by different people or using different methods, are highly correlated, but principals were found to make the lowest ones. Teachers' estimates and the application of district allocation formulas were found to be very close to the estimates based on the federal allocation formula applied to income data from home interviews, which we assumed to be the most accurate. No systematic bias in estimates was found to be related to such school characteristics as urbanism, enrollment, or minority concentration. Principals were found to under estimate school poverty, but they do it so consistently that district rankings of schools on that basis can still be expected to conform well to Title I intentions.

The third chapter is closely related insofar as it provides survey data on the criteria that districts use to assess school poverty en route to the determination of Title I eligibility. Over 78 percent of the nation's school districts are projected to use counts of free or reduced-price lunches, and about 60 percent use AFDC enrollments. Most districts use more than one criterion, but we can't tell if they use them jointly in order to converge on 'true' estimates of school poverty or to select schools on the basis of other considerations not within the intent of the Title I law.

The last two chapters provide information on where Title I services are being distributed. We provide tabulations of Title I services by school poverty, average school achievement, minority concentration, and enrollment. As would be expected, poorer schools have higher rates of Title I participation than less poor schools, but the relationship is not perfect. The Title I participation rates by school poverty level are presented below by region and urbanism.

Region	Percentage of Lower Poverty Schools Participating in Title I	Percentage of Higher Poverty Schools Participating in Title I
New England	52	79
Metropolitan Northeast	61	84
Mid-Atlantic	56	89
Southeast	37	89
North Midwest	61	79
South Central	39	88
Central Midwest	64	82
North Central	42	75
Pacific Southwest	14	69
Pacific Northwest	50	96

Urbanism	Percentage of Lower Poverty Schools Participating in Title I	Percentage of Higher Poverty Schools Participating in Title I
City of over 500,000	14	74
City of 200,000-500,000	13	73
City of 50,000-200,000	18	79
Suburb of city	46	74
City under 50,000	57	91
Rural area near city	68	89
Rural area not near city	66	88

Considering only schools participating in Title I, the average achievement percentile scores for three different kinds of students are presented below.

Participation of Student	Average Reading Percentile	Average Math Percentile
Title I students	26.5	28.4
Other-CE students	37.1	40.1
Non-CE students	51.8	49.0

The trend is quite clear that the lowest achieving students are selected for Title I services.

Minority children participate in Title I programs in higher proportions than their incidence in the population, mostly because they generally attend schools in high-poverty areas and have low achievement levels. Minority participation is highest in the high-poverty schools and in the cities.

We also briefly examine certain dimensions of the educational services received in the name of Title I, and how students are selected to receive them. Title I participants are offered the same number of days of instruction per school year, but have slightly higher absence rates. About two-thirds of the CE students continue to receive CE for more than one year. While CE students are receiving their CE services, most of their non-CE peers are receiving instruction in the same subject, but presumably not of a compensatory or remedial nature.

Finally, we tabulate the incidence of Title I services and services for special students in order to disentangle the two largest federal education programs, Title I and P.L. 94-142, in terms of their services and effects. The approximate percentages of students with physical, psychological, or adjustment problems that are judged to interfere with academic performance by CE status are provided below.

CE Status of Students	Percentage of Students Judged to Have Problems
Title I only	18
Title I and Other CE	23
Other CE only	20
No CE	9

ACHIEVEMENT AND COMPENSATORY EDUCATION

Part II contains two chapters that are concerned with how and how well Title I and CE in general are distributed at the student level. In the first chapter, we develop several indexes of school-level 'targeting'—the extent to which low achievers are selected for CE participation The application of any of the targeting indexes to all schools, without regard to the philosophies underlying their CE programs and the wide range of acceptable structuring of the programs, results in some unfair verdicts. The indexes are compared on the basis of their relative merits, but none can be universally recommended.

The second chapter relates the results of scores on achievement tests to teachers' judgments of students' needs for CE. We wanted to determine how much agreement there was so that, if agreement were high, the less costly judgments could be encouraged as acceptable methods for selecting CE participants. The correlation between scores and judgments is about 0.5, overall. If achievement scores are dichotomized near the 35th percentile, agreement with teacher judgments is generally maximized, with over 75 percent of the students categorized consistently as needing or not needing CE.

Judgments being what they are, we also studied what might influence them so that, even if accurate, systematic biases or errors could be pointed out. Although such biases are evident, they are of very small magnitude.

MEASURING STUDENT GROWTH

In any study that is meant to influence social policy, the measures employed to assess outcomes of the programs are critical. In the case of the SES, the outcomes are the growth rates of groups of students who do and do not participate in CE programs. Part III describes the procedures we employed in the selection and development of the student-level measures of academic growth, practical-achievement, and affect. We also investigate issues of testing that have been raised by the SES.

The measures of academic growth, which are the subject of the first two chapters of Part III, are standardized tests and criterion-referenced tests of reading and math. Skills in reading and math underlie academic achievement in all other areas and are, therefore, the objectives of most CE programs. After a careful evaluation of all standardized achievement tests, the Comprehensive Tests of Basic Skills, Form S, was selected. This test was administered at all grades in both the fall and spring. During the first year of the study, each student was tested at the appropriate level and at the next-lower level. The double testing allowed for the construction of an accurate vertical scale of growth which, in turn, made it possible to test at a single functional level during the later years of the study. The test series also underwent several analyses in an effort to identify items biased against minority students. Items identified as biased were not scored for any of the analyses of the study.

The measurement needs of the study and the theoretical underpinnings and practical limitations of criterion-referenced testing were also assessed to determine the feasibility of employing such an approach in the nationwide study. In general, the practical limitations of criterion-referenced tests and the level of effort needed to remedy those limitations forced us to conclude that the use of such instruments was not feasible.

When the very value of academic skills is called into question, it usually is with the suspicion that such skills do not have much functional value for many individuals. The result of this line of thinking has been an increased interest in measures of what has been called 'functional literacy.' The concept of functional literacy, treated in the third chapter, has been expanded in the SES to include functional math skills and thus is called practical achievement. Upon completion of a review of theory and practice in the assessment of functional literacy and practical achievement, it was concluded that no existing instrument would be appropriate for the students in the SES. Consequently, an instrument-development effort was undertaken that resulted in a reliable and student-acceptable instrument that taps practical reading and computing.

The value of affective growth can be justified either as a goal in itself or as an instrumental goal necessary for academic growth. Either way, measures of affective growth have occupied an important position in the batteries of outcome measures developed for the study of educational programs. These measures are addressed in the fourth chapter of Part III Available measures of aspects of affective growth in children were reviewed with the intent of selecting a set of scales as outcome-measures. The Survey of Student Attitudes was selected to assess attitudes toward reading and math. Additional items were generated in a similar format to assess attitudes toward school in general. These three dimensions of attitude appeared in the field test to be relatively independently and reliably measured constructs.

In the fifth chapter we look at 'out-of-level' testing—a practice whereby very low-achieving students are tested with a level of test designed for students in a lower grade. Average school percentages of scores below the chance level and at the ceiling obtained with at-and below-level tests are presented below.

	Average School Percentages of Students Scoring at the Floor and Ceiling			
Test Level	At- or Below-Chance Scores		At- or Near-Perfect Scores	
	Reading	Math	Reading	Math
Grade-Level	19.2	17.6	0 8	0.1
Below-Level	5.6	2.9	8.3	4.7

Testing with below-level tests greatly reduces 'floor' effects, but increases 'ceiling' effects somewhat. Out-of-level testing has much to recommend it on several grounds, but it also presents problems. The pros and cons are discussed in some detail in this chapter.

In the last chapter we investigate the issue of the 'speededness' of the achievement tests used in the SES. Our special concern is to determine if the speed factor influences scores differently for different racial/ethnic groups. In the event that such effects can be uncovered, and speed is considered an irrelevant component of achievement, testing procedures in the SES should be modified to eliminate speed so that research findings are not influenced by it. We found that close to 90 percent of the students attempt the last item of regularly timed tests. Test speededness, as a phenomenon, is found to be most pronounced for black students at the higher grades and less pronounced for brown students.

THE EFFECTS OF ATTRITION

Part IV is composed of two chapters that address two distinct issues of attrition from the longitudinal study. In the first, we describe the reduction in the SES sample of schools in the second year of the study that resulted from funding cutbacks. Although the retained sample was selected with specific study purposes in mind, and could not be assumed to be random and representative, tests of the data show its remarkable ability to approximate the population and the first-year sample. We concluded that the reduced sample will serve the SES well, but that the reduction may be expected to have introduced some complications for the interpretation of some analyses.

Based on the first full calendar year of the SES, we document in the second chapter the incidence of attrition of students and compare rates over several characteristics, such as achievement level and minority status. The average achievement scores for three groups of students are given below.

Type of Student	Average Reading Percentile	Average Math Percentile
Students who stay in the study	49.7	50.0
Students who left the study	40.8	42.3
Students who enter the study late	44.7	44.5

We can see that the sample of stable students has a higher achievement level than either attrital group. When minority status and poverty are calculated for the three groups, the findings are:

Type of Student	Percent That Are Minority	Percent That Receive Free or Reduced-Price Meals
Students who stay in the study	30.8	39.0
Students who left the study	36.9	45.6
Students who enter the study late	34.0	50.5

In addition to being lower achievers, students in the attrition groups are more likely to be minority students and to receive free or reduced-price meals. We also study the attrition rates by several characteristics jointly, in order to understand better the major sources of attrition. Finally, we provide some conjectures about the expectable influences of the observed attrition on different kinds of analyses and issues in the SES.

PART I. ALLOCATION OF TITLE I FUNDS AND SERVICES

Part I contains five chapters that are collections of analyses that were especially requested by the Department of Education to provide information for the Congressional hearings on the reauthorization of Title I in 1978. The first chapter addresses poverty concentrations in districts and schools in order to provide estimates of the potential incidence of districts and schools with sufficiently high concentrations of poor children so that they can automatically qualify for special eligibility for Title I services. We use a modeling-like process to make projections of the national incidence of such districts and schools under different poverty-concentration criteria to determine the distribution of such districts and schools over regions and categories of urbanism. In the second chapter we study the results of different methods for defining school poverty in the hope not only of checking the validity of such information, but also to insure that weaknesses of method can be pinpointed and advertised, thus discouraging such methods from common use. The third chapter is closely related insofar as it provides survey data on the criteria that districts currently use to assess the poverty level of schools.

The last two chapters provide information on where Title I services are being provided. We provide tabulations of Title I services by school poverty, school achievement, minority concentration, and enrollment. We also take a brief look at certain dimensions of those services, as well as how students are selected to receive them. Finally, we tabulate the incidence of Title I services as opposed to services for special (handicapped) students in order to disentangle the services and effects of the two largest federal education programs (ESEA Title I and P.L. 94-142). Thus, findings of the Sustaining Effects Study will apply more directly to compensatory-education efforts and not to special-education efforts.

CHAPTER 1. SCHOOLS AND DISTRICTS QUALIFYING FOR TITLE I UNDER DIFFERENT CRITERIA OF POVERTY CONCENTRATION

Ralph Hoepfner
Ming-mei Wang
Moraye B. Bear
Henry J. Zagorski

Data from the Sustaining Effects Study (SES) were projected to the nation in order to supply the U.S. Department of Education with estimates of the number of districts, schools, and students affected by changes in the 1978 amendments to Title I, especially those changes affecting various eligibility criteria that incorporate the notion of the 'concentration' of poor children. Alternative definitions of concentration were used for the projections so that Congress would be informed of likely consequences of the changes. When poverty concentrations are considered, either to establish target areas, to enable school-wide projects, or to allocate concentration grants to districts, the results of the uneven national distribution of wealth can be seen. The Southeast and South Central regions, and large cities and rural areas, where poverty is more common, tend to be the beneficiaries of policies that consider concentrations of poverty in addition to the considerations already in the Title I allocation formulas and eligibility rules.

Although the reauthorization of Title I in 1978 did not contain major changes in the distribution formulas for funds, certain minor adjustments were made in an attempt to improve conformance of the program to its underlying intention to increase the equality of educational opportunity. Among these adjustments were clarifications about what have become known as 'concentration grants' and 'target areas.'

Each year, the Title I allocation is distributed to school districts by way of three types of grants. Basic Grants are those sub-allocations distributed according to a complicated, but uniform procedure to all eligible districts. Because the imposition of a uniform procedure cannot meet all of the intentions of the Congress, two additional types of grants are also made. Incentive Grants are awarded to states that do better than the national average in supporting compensatory education. Through Incentive Grants, the federal government encourages state and local education programs by rewarding those states that financially demonstrate their commitments to compensatory education. The third type of grant is the Concentration Grant, made to districts with either large numbers or high proportions of poor children. These grants addresss the belief that educational problems are compounded in areas with relatively high concentrations of poverty. The proportion of poor children that qualifies a district to receive a Concentration Grant was set in 1978 at 20 percent (or 5,000 students), but in the process of setting this proportion it was important to examine the results of various proportions in terms of the numbers and types of districts that would qualify.

At the school level, the concentration of poor children is almost important, because within districts, school eligibility for Title I is determined by the concentration of poverty in the school's attendance area. But even further, among schools eligible under Basic Grants, those with certain concentrations of poverty can implement school-wide Title I projects in which all the students in the school can be served instead of only the lowest-achieving ones, when certain fiscal requirements are also met. When schools have very high proportions of students from low-income families (set in the Education Amendments of 1978 at 75 percent), then after some local or state funds are added, all the children may be served by a school-wide program. In a similar vein, the concept of 'target area' establishes a uniform baseline for the automatic eligibility of all schools with a certain concentration of poor children (set at 20 percent in the Education Amendments of 1978). Implementation of this rule means that no school at or above the concentration level can be ineligible for Title I due to its low poverty ranking within its district, while another school with a lower concentration is eligible because it ranked high in its (higher income) district. This can be seen as the imposition of an absolute standard onto the relative standards resulting from the distribution formula's sensitivity to the nation's complex economic and political structure.

In summary, in terms of districts with concentrations of poor children, Concentration Grants would be awarded, while in terms of schools with high concentrations, eligibility is assured and all students can be served. These exceptions to the usual allocations (via Basic Grants) are justified both on practical grounds (it's difficult not to serve a small number of students in a school), and on educational grounds (the high concentration of deprived students is presumed to have deleterious effects on the academic growth of all the students).

Prior to and during the Congressional hearings on reauthorization, the U.S. Department of Education requested pertinent information from the SES regarding how many districts, schools, and students would be affected under implementation of the rules with different criteria for 'concentration.' The value of the SES data lay not only in its high degree of accuracy when statistically weighted for the population (see Hoepfner, Zagorski, and Wellisch, 1977), but also in its recency (collected in 1976 and 1977). The only alternative to these recent data were data from the 1970 census, considered to be inadequate due to subsequent demographic changes. The information was to be used in the following manner. for various definitions of 'concentration,' we would estimate the numbers of districts, schools, or students qualifying, and describe them in terms of region and urbanism.

SCHOOLS IN TARGET AREAS

In 1975-76, 4,750 principals in a nationally representative sample were queried about the characteristics of their schools. In particular, they responded to one item that asked for an estimate of the percentage of students meeting Title I poverty criteria (see the third chapter in this part) or the percentage of students eligible for free lunches. When the principal provided both percentages, the higher one was chosen as an estimate of the percentage of the school's enrollment from homes in poverty (the school's poverty concentration). The grade 1-6 enrollment, geographic location, and degree of urbanism were also obtained from responses to the questionnaire. What follows are results of our tabulation for five definitions of 'concentration' counts. For each we present the numbers and types of schools qualifying.

Schools With Ten Percent or More Poverty

We project that there are 47,146 schools offering any of the grades in the 1-6 range that have 10 percent or more of students in poverty. This amounts to 75.4 percent of the nation's elementary schools. Clearly then, if 10 percent were specified as the concentration of poor children ensuring school eligibility for Title I, then over 75 percent of the elementary schools would be automatically eligible. Given reasonable estimates of the future funding for Title I, the implications are that either many schools would be defined as eligible but not be able to participate, or the program would have to be spread out to so many schools that the per-pupil allocation would be grossly insufficient for provision of programs that are of adequate size, scope, and quality.

Regional Distribution of the 47,146 Schools. Table 1-1 presents the projected distribution of the '10 percent poverty' schools among ten regions. With the exceptions of the Southeast, North Midwest, and South Central regions the projected percentages of schools having 10 percent or more of poverty enrollments correspond closely to the percentages of elementary schools in the regions. A disproportionately greater number of schools have 10 percent or more poor children enrolled in the Southeast and South Central regions, disproportionately fewer schools in the North Midwest region.

Table 1-1

Projected Regional Distribution of Elementary Schools With Poverty Concentrations at and Above 10 Percent

	Schools With 10 Percent or More Poverty		
Region*	**Projected Number**	**Projected Percentage**	**Population Percentage of Elementary Schools With Grades 1-6**
New England	3,444	7.3	7.1
Metropolitan Northeast	3,351	7.1	8.3
Mid-Atlantic	4,884	10.4	10.6
Southeast	8,668	18.4	14.6
North Midwest	8,606	18.3	21.9
South Central	6,161	13.1	11.1
Central Midwest	3,437	7.3	7.1
North Central	2,257	4.8	5.2
Pacific Southwest	4,354	9.2	9.9
Pacific Northwest	1,987	4.2	4.2
Total	47,149**	100.1**	100.0

*The regions are New England - CN, MA, ME, NH, RI, VT, Metropolitan Northeast - NJ, NY, Mid-Atlantic - DE, DC, MD, PA, VA, WV, Southeast - AL, FL, GA, KY, MS, NC, SC, TN, North Midwest - IL, IN, MI, MN, OH, WI; South Central - AR, LA, NM, OK, TX, Central Midwest - IA, KN, MO, NE, North Central - CO, MT, ND, SD, UT, WY; Pacific Southwest - AZ, CA, HI, NV; Pacific Northwest - AK, ID, OR, WA.

**Discrepancies are due to rounding errors 47,149 is 75.4 percent nationwide

Distribution of the 47,146 Schools by Urbanism. Table 1-2 presents the distribution of the '10 percent poverty' schools among seven categories of urbanism. Comparing the last two columns of Table 1-2, it can be seen that suburbs of cities have a smaller projected percentage of schools with 10 percent or more poverty than its share of schools in the population, while rural areas not near cities have larger projected percentages. Defining target areas in terms of 10 percent concentrations of low-income children, therefore, offers potential benefits more to rural areas than to the suburbs. Cities would also experience some small benefits under this criterion.

Table 1-2

Projected Distribution of Elementary Schools With Poverty Concentrations at and Above 10 Percent, by Categories of Urbanism

Schools With 10 Percent or More Poverty			
Urbanism	Projected Number	Projected Percentage	Projected Population Percentage of Elementary Schools With Grades 1-6
City of over 500,000	3,909	8 3	7 3
City of 200,000-500,000	2,033	4 3	3 9
City of 50,000-200,000	5,084	10 8	10.6
Suburb of city	4,285	9 1	14 8
City under 50,000	13,285	28 2	28.0
Rural area near city	4,969	10.5	10.6
Rural area not near city	13,585	28 8	24.9
Total	47,150*	100 0	100 1

*Discrepancies are due to rounding errors, 47,150 is 75 4 percent nationwide

Schools With Twenty Percent or More Poverty

Because of the large number of the nation's elementary schools that would be eligible for Title I with 10 percent concentrations of students from low-income families, similar tabulations were made with a 20 percent criterion. The national projection indicates that 33,337 (53.3 percent) of the nation's elementary schools with grades in the 1-6 range have 20 percent or more of their students from low-income families. By subtraction, using the figures from the first criterion above, we can see that 13,809—21.1 percent—have poverty concentrations between 10 and 19 percent. Clearly, even this criterion allows more schools to be eligible to participate in Title I than are likely to be well served under reasonable expectations of future funding for the program.

Regional Distribution of the 33,337 Schools. Table 1-3 presents the distribution of the 33,337 schools with 20 percent concentrations of students in poverty across the ten regions of the nation. Comparisons of the last two columns reveal that the Southeast and South Central regions have a higher proportion of schools with 20 percent or more poverty enrollments than their population share of schools. Conversely, the North Midwest region has few schools with 20 percent or more poverty enrollments in comparison to its population share of schools. The distribution does not differ much from that found for schools with 10 percent or more poverty, but a noticeable difference is that the percentage becomes larger for the Southeast (18.4 percent in the 10 percent or more poverty case, compared to 22.9 percent for the 20 percent case), while it decreases for the North Midwest (from 18.3 percent to 15.0 percent). The Southeast and South Central regions clearly have an advantage under the 20 percent criterion in terms of gaining eligibility for schools, while the North Midwest and, to a lesser extent, the Central Midwest, have a disadvantage.

Table 1-3

Projected Regional Distribution of Elementary Schools With Poverty Concentrations at and Above 20 Percent

	Schools With 20 Percent or More Poverty		
Region	Projected Number	Projected Percentage	Population Percentage of Elementary Schools With Grades 1-6
New England	2,243	6 7	7 1
Metropolitan Northeast	2,220	6 7	8.3
Mid-Atlantic	3,541	10 6	10 6
Southeast	7,642	22 9	14.6
North Midwest	4,992	15.0	21.9
South Central	5,198	15.6	11.1
Central Midwest	1,922	5 8	7.1
North Central	1,276	3 8	5 2
Pacific Southwest	3,110	9 3	9 9
Pacific Northwest	1,196	3.6	4.2
Total	33,340*	100 0	100.0

*Discrepancies are due to rounding errors, 33,340 is 53 3 percent nationwide.

Distribution of the 33,337 Schools by Urbanism. The projected distribution of the 33,337 schools with 20 percent or more poverty enrollments over the seven categories of urbanism is presented in Table 1-4. By comparing the last two columns, it is seen that large cities and rural areas not near cities have greater percentages of schools with 20 percent or more poverty enrollments than their population percentages. In contrast, suburbs have considerably smaller percentages than their population percentages of elementary schools. Not surprisingly, the distribution is quite similar to that of schools with 10 percent or more poverty.

Table 1-4

Projected Distribution of Elementary Schools With Poverty Concentrations at and Above 20 Percent, by Categories of Urbanism

	Schools With 20 Percent or More Poverty		
Urbanism	Projected Number	Projected Percentage	Projected Population Percentage of Elementary Schools With Grades 1-6
City of over 500,000	3,549	10.6	7.3
City of 200,000-500,000	1,627	4.9	3.9
City of 50,000-200,000	3,690	11.1	10.6
Suburb of city	2,143	6.4	14.8
City under 50,000	9,112	27.3	28.0
Rural area near city	2,950	8.8	10.6
Rural area not near city	10,268	30.8	24.9
Total	33,339*	99.9	100.1

*Discrepancies are due to rounding errors; 33,339 is 53.3 percent nationwide.

Schools With Thirty Percent or More Poverty

Our projections estimate that 22,973 (36.7 percent) of the nation's elementary schools with grades in the 1-6 range have 30 percent or more of their enrollments from poverty backgrounds. We can infer from these and earlier figures that 10,364 schools—16.6 percent of the nation's elementary schools—have between 20 to 29 percent poverty enrollments.

Regional Distribution of the 22,973 Schools. Inspection of the last two columns of Table 1-5 indicates that the Southeast and South Central regions have greater percentages of schools with 30 percent or more poverty enrollments in comparison with their population percentages. The table also reveals that the percentage of schools with 30 percent or more poverty enrollments in the North Midwest is just slightly more than half of its population of elementary schools. Again, we can observe the tendency of an increasing share (from 22.9 percent in the case of 20 percent or more poverty, to 27.9 percent here) of schools meeting the poverty criterion in the Southeast region as the cutoff percentage of poverty enrollments is raised. On the other hand, there is a decreasing share (from 15.0 percent in the case of the 20 percent cutoff to 12.6 percent when the cutoff is 30 percent) in the North Midwest region.

Distribution of the 22,973 Schools by Urbanism. The last two columns of Table 1-6 reveal that the percentages of schools with 30 percent or more poverty enrollments are larger for large cities (over 500,000) and rural areas not near cities in comparison to their population percentages of elementary schools with grades in the 1-6 range. As one raises the school poverty criterion from 10 percent to 20 percent to 30 percent, large cities gain in their shares of schools in target areas. In contrast, the share of schools meeting the poverty criterion in suburbs decreases from 9.1 percent to 6.4 percent to 4.7 percent.

Table 1-5

Projected Regional Distribution of Elementary Schools With Poverty Concentrations at and Above 30 Percent

Schools With 30 Percent or More Poverty			
Region	**Projected Number**	**Projected Percentage**	**Population Percentage of Elementary Schools With Grades 1-6**
New England	1,450	6.3	7.1
Metropolitan Northeast	1,517	6.6	8.3
Mid-Atlantic	2,503	10.9	10.6
Southeast	6,409	27.9	14.6
North Midwest	2,902	12.6	21.9
South Central	4,002	17.4	11.1
Central Midwest	997	4.3	7.1
North Central	639	2.8	5.2
Pacific Southwest	2,009	8.7	9.9
Pacific Northwest	547	2.4	4.2
Total	22,975*	99.9	100.0

*Discrepancies are due to rounding errors; 22,975 is 36.7 percent nationwide.

Table 1-6

Projected Distribution of Elementary Schools With Poverty Concentrations at and Above 30 Percent, by Categories of Urbanism

Schools With 30 Percent or More Poverty			
Urbanism	**Projected Number**	**Projected Percentage**	**Projected Population Percentage of Elementary Schools With Grades 1-6**
City of over 500,000	3,097	13.5	7.3
City of 200,000-500,000	1,244	5.4	3.9
City of 50,000-200,000	2,836	12.3	10.6
Suburb of city	1,076	4.7	14.8
City under 50,000	5,858	25.5	28.0
Rural area near city	1,676	7.3	10.6
Rural area not near city	7,188	31.3	24.9
Total	22,975*	100.0	100.1

*Discrepancies are due to rounding errors; 22,975 is 36.7 percent nationwide.

Comparisons Among Poverty Percents

Figures 1-1 and 1-2 present the data from Tables 1-1, 1-3 and 1-5 and from Tables 1-2, 1-4, and 1-6, respectively, in graphic form for ease of seeing the effects by region and urbanism of the various criteria for target areas. With higher and higher criteria, the Southeast and South Central regions and large cities and rural areas gain in the percentages of eligible schools. At the same time, suburbs and the North Midwest lose.

Schools With Poverty Enrollments of 50 or More Students

So far, we have limited our concern with poverty concentrations to criteria expressed as percentages of school enrollment. Percentages based on small enrollments may have different educational implications from those based on large enrollments, so we did similar tabulations using an absolute criterion of 50 or more low-income students. Use of an absolute number like 50 provides a less relative assessment of the potential impact of poverty, and would be expected to reduce the rate of eligibility in sparsely-populated areas that have small schools. We project that 31,743, or 50.8 percent of the elementary schools with grades in the 1-6 range have 50 or more low-income students.

Regional Distribution of the 31,742 Schools. Inspection of Table 1-7 reveals that the Southeast has the largest number of schools with poverty enrollments of 50 or more students; the North Midwest has the smallest number of such students.

Table 1-7

Projected Regional Distribution of Elementary Schools With 50 or More Low-Income Students

	Schools With 50 or More Poor Students		
Region	**Projected Number**	**Projected Percentage**	**Population Percentage of Elementary Schools With Grades 1-6**
New England	1,688	5.3	7.1
Metropolitan Northeast	2,714	8.5	8.3
Mid-Atlantic	3,521	11.1	10.6
Southeast	7,915	24.9	14.6
North Midwest	4,928	15.5	21.9
South Central	4,636	14 6	11.1
Central Midwest	1,266	4.0	7.1
North Central	815	2.6	5.2
Pacific Southwest	3,254	10.3	9.9
Pacific Northwest	1,008	3.2	4.2
Total	31,745*	100.0	100.0

*Discrepancies are due to rounding errors; 31,745 is 50 8 percent nationwide.

44

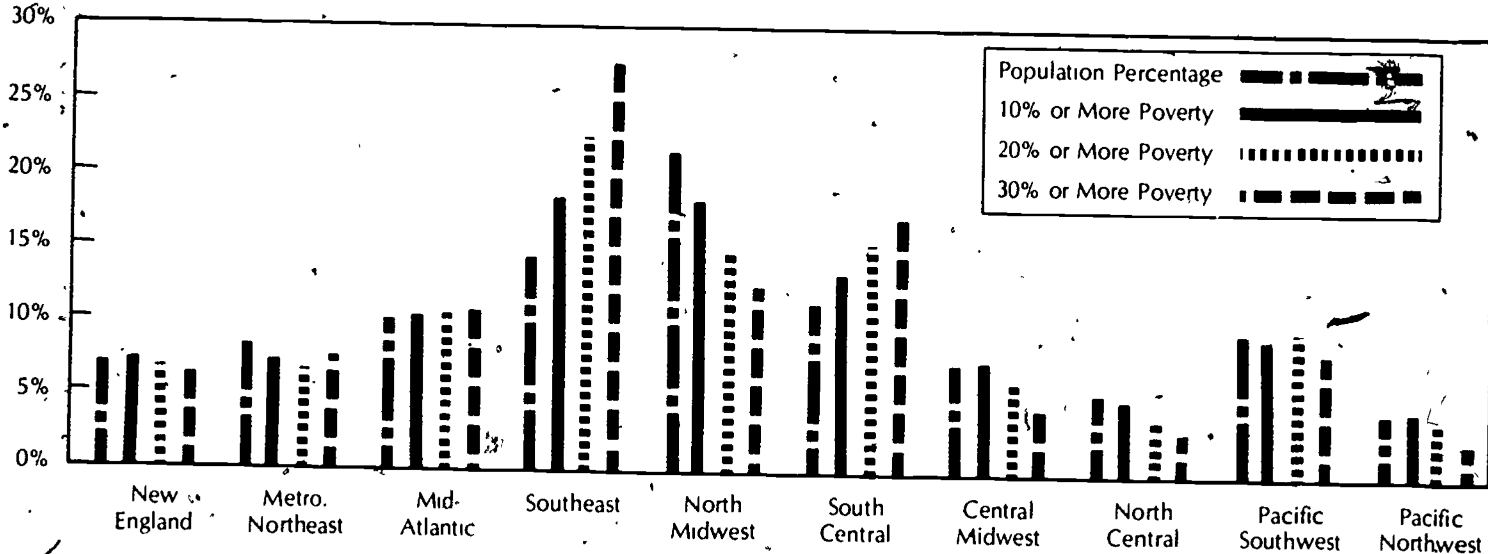

Figure 1-1

Percentages of the Nation's Elementary School Students in Each Region Attending Schools Eligible Under Each of Three Criteria

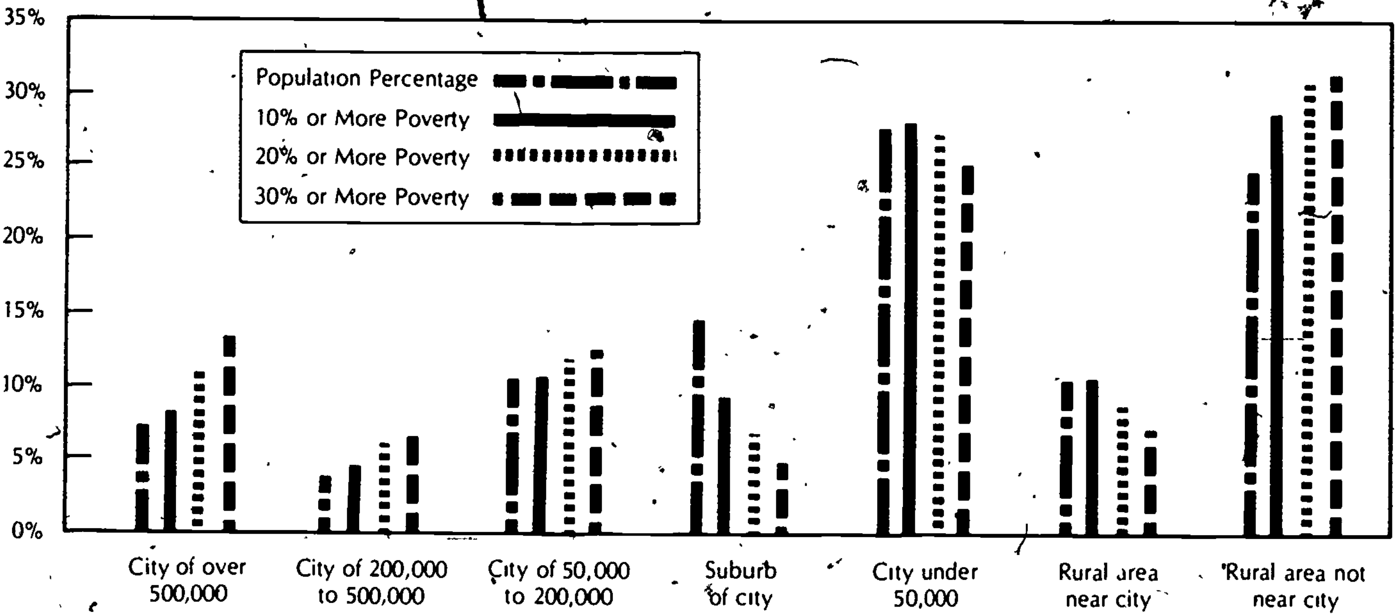

Figure 1-2

Percentages of the Nation's Elementary School Students, by Urbanism, Attending Schools Eligible Under Each of Three Criteria.

Distribution of the 31,743 Schools by Urbanism. Table 1-8 shows that urban areas (the first three categories within urbanism) have the largest numbers of schools with 50 or more students in poverty. Rural areas do not suffer a large disadvantage with this criterion, but the suburbs do.

Table 1-8

Projected Distribution of Elementary Schools With 50 or More Low-Income Students, by Categories of Urbanism

Urbanism	Projected Number	Projected Percentage	Projected Population Percentage of Elementary Schools With Grades 1-6
	Schools With 50 or More Poor Students		
City of over 500,000	3,769	11.9	7.3
City of 200,000-500,000	1,734	5.5	3.9
City of 50,000-200,000	4,050	12.8	10.6
Suburb of city	3,142	9.9	14.8
City under 50,000	9,078	28.6	28.0
Rural area near city	2,857	9.0	10.6
Rural area not near city	7,115	22.4	24.9
Total	31,745*	100.1	100.1

*Discrepancies are due to rounding errors; 31,745 is 50.8 percent nationwide.

Schools With 10 Percent or More Poverty or 50 or More Low-Income Students

Finally, we present the number of schools in the nation with *either* 10 percent or more poverty *or* 50 or more students from low-income families. We estimate in Table 1-9 that 47,397 (75.8 percent) of the nation's elementary schools with grades in the 1-6 range have either 50 or more students in poverty or 10 percent or more of their 1-6 enrollments in poverty. Note that the use of the 10 percent criterion alone yielded 75.4 percent; so the addition of another possibility—50 or more students—did not add many schools. More than half of the schools meet both of the criteria if they meet one.

Table 1-9

Projected Numbers and Percentages of Elementary Schools With Either 10 Percent or More Poverty or 50 or More Low-Income Students

	9% or Less Poverty	10% or More Poverty	Total
49 or fewer low-income students	15,130 24.2%	15,655 25.0%	30,785* 49.2%
50 or more low-income students	251 0.4%	31,491 50.4%	31,742* 50.8%
Total	15,381 24.6%	47,146 75.4%	62,527 100.0%

*Discrepancies are due to rounding errors

DISTRICTS ELIGIBLE FOR CONCENTRATION GRANTS

The discussion so far has focused on the characteristics of schools for determining their eligibility to offer Title I services. The concept of poverty concentrations is also relevant to districts for determining their eligibility to receive a Title I Concentration Grant. We therefore used the SES representative sample of districts to project eligibility rates for Concentration Grants.

Table 1-10 presents national estimates by tenths of district-poverty concentrations. The projections are based on 219 superintendents' estimates of the percentages of students in their districts that met the Title I poverty criteria during the 1976-77 school year. This sample was projected to 14,683 districts that served children in the grade 1-6 range.

If additional funds were allocated to districts with 20 percent or more enrollments of children from low-income families according to these 1976-77 data, we estimate that 48 percent of the nation's districts* would qualify for them. About one-fourth of the nation's students are enrolled in schools in those districts.

Table 1-10

Projected Distribution of the Nation's School Districts, by Percentage of Students in Poverty

Percentage of Students in Poverty	National Projections From 219 Districts in the SES	
	Number of Districts	Percentage of Districts
90-100	113	0.8
80-89	15	0.1
70-79	305	2.1
60-69	136	0.9
50-59	928	6.3
40-49	1,695	11.5
30-39	1,023	7.0
20-29	2,828	19.3
10-19	5,178	35.3
0-9	2,462	16.8
Total	14,683	100.1

*It is interesting to note the data used for determining district eligibility for Concentration Grants under Section 117 of Title I result in about half the nation's districts receiving such grants. About 80 percent of them are in urban settings.

SCHOOLS ELIGIBLE FOR SCHOOL-WIDE PROJECTS

At a certain saturation of a school's enrollment with low-income children, it seems practically and educationally unnecessary to go through the trouble of assessing student need and then selecting among them for Title I services. The chances are presumed high in these cases that all students are in need and the school's Title I resources could be (beneficially) distributed to all of them. This is the rationale behind the 1977-78 discussion about school-wide projects for the Education Amendments. To inform this discussion, we were asked to provide the Office of Education with national projections of schools with 70 percent and 80 percent enrollments of low-income children.

Schools With 70 Percent or More Poverty. We estimate that 9.2 percent (5,724) of the nation's schools with grades in the 1-6 range have enrollments of 70 percent or more of children from low-income families. These schools serve an estimated 2,150,279 students (about 10.3 percent of the nation's public school students enrolled in grades 1-6). The difference in the school and student percentages indicates that the high-poverty schools tend to be larger than average.

Tables 1-11 and 1-12 present the distribution of these schools by region and by urbanism. From the tables we can see that the Southeast and South Central regions have an abundance of these high-poverty schools, while the North Midwest has relatively few of them. Similarly, large cities could have more school-wide projects under this criterion and suburbs and small cities could have fewer of them.

Table 1-11

Projected Regional Distribution of Elementary Schools With Poverty Concentrations at and Above 70 Percent

Schools With 70 Percent or More Poverty			
Region	Projected Number	Projected Percentage	Population Percentage of Elementary Schools With Grades 1-6
New England	425	7.4	7.1
Metropolitan Northeast	457	8.0	8.3
Mid-Atlantic	513	9.0	10.6
Southeast	1,620	28.3	14.6
North Midwest	590	10.3	21.9
South Central	1,172	20.5	11.1
Central Midwest	176	3.1	7.1
North Central	124	2.2	5.2
Pacific Southwest	528	9.2	9.9
Pacific Northwest	120	2.1	4.2
Total	5,725*	100.1	100.0

*Discrepancies are due to rounding errors; 5,725 is 9.2 percent nationwide.

Table 1-12

Projected Distribution of Elementary Schools With Poverty Concentrations at and Above 70 Percent, by Categories of Urbanism

Schools With 70 Percent or More Poverty			
Urbanism	**Projected Number**	**Projected Percentage**	**Projected Population Percentage of Elementary Schools With Grades 1-6**
City of over 500,000	1,674	29.2	7.3
City of 200,000-500,000	426	7.4	3.9
City of 50,000-200,000	633	11.1	10.6
Suburb of city	88	1.5	14.8
City under 50,000	1,019	17.8	28.0
Rural area near city	208	3.6	10.6
Rural area not near city	1,676	29.3	24.9
Total	5,724*	99.9	100.1

*Discrepancies are due to rounding errors; 5,724 is 9.2 percent nationwide.

Schools with 80 Percent or More Poverty. When the criterion for school-wide projects is 80 percent, the percentage of qualifying schools is 5.8 (3,638 schools). We project that 1,383,008 students (about 6.6 percent) are enrolled in the schools with 80 percent or more of students from low-income families. Tables 1-13 and 1-14 present the distributions of these high-poverty schools over region and urbanism.

Table 1-13

Projected Regional Distribution of Elementary Schools With Poverty Concentrations at and Above 80 Percent

Schools With 80 Percent or More Poverty			
Region	**Projected Number**	**Projected Percentage**	**Population Percentage of Elementary Schools With Grades 1-6**
New England	192	5.3	7.1
Metropolitan Northeast	329	9.1	8.3
Mid-Atlantic	233	6.4	10.6
Southeast	1,033	28.4	14.6
North Midwest	442	12.1	21.9
South Central	785	21.6	11.1
Central Midwest	105	2.9	7.1
North Central	102	2.8	5.2
Pacific Southwest	318	8.7	9.9
Pacific Northwest	98	2.7	4.2
Total	3,637*	100.0	100.0

*Discrepancies are due to rounding errors; 3,637 schools is 5.8 percent nationwide.

Table 1-14

Projected Distribution of Elementary Schools With Poverty Concentrations at and Above 80 Percent, by Categories of Urbanism

Urbanism	Schools With 80 Percent or More Poverty: Projected Number	Projected Percentage	Projected Population Percentage of Elementary Schools With Grades 1-6
City of over 500,000	1,282	35.2	7.3
City of 200,000-500,000	276	7.6	3.9
City of 50,000-200,000	367	10.1	10.6
Suburb of city	43	1.2	14.8
City under 50,000	510	14.0	28.0
Rural area near city	75	2.1	10.6
Rural area not near city	1,086	29.8	24.9
Total	3,639*	100.0	100.1

*Discrepancies are due to rounding errors; 3,639 schools is 5.8 percent nationwide.

Based in part on these projections, an intermediate level (75 percent) was specified in the amendment law.

In summary, when the concentration of children from low-income families is considered, either to establish target areas, to enable school-wide projects, or to allocate concentration grants to districts, the results of the uneven national distribution of wealth can be seen. The Southeast and South Central regions and large cities and rural areas, where poverty is more common, tend to be the beneficiaries of policies that consider concentrations of poverty in addition to the consideration already in the Title I allocation formulas.

CHAPTER 2. ESTIMATES OF SCHOOL POVERTY

Joel M. Moskowitz and Deborah S. Brown

Districts are allowed to use a variety of ways to estimate the number or proportion of poor children in schools for the purpose of determining the eligibility of schools to offer Title I services. It is important, therefore, to know how the different ways result in estimates of different accuracy. Five estimates of the concentration of poor students in schools were developed and tested. Although the estimates are moderately to highly intercorrelated, they result in significantly different average estimates. Using the Federal Allocation Formula applied to income data obtained from home interviews as the standard and accurate measure of school poverty, estimates by principals were found to be significantly and consistently lower; and estimates by teachers and by application of district allocation formulas to be very close to the standard. The most commonly used estimates of school poverty levels, principals' reports, while resulting in significantly lower absolute estimates, are so consistent in their underestimation that districts' ordering of schools can be expected to conform well to the Title I intentions.

The distribution of Title I money from the federal level through the states, counties, districts, and schools—ending up as additional services for specific kinds of students—is a complex procedure. In essence, the flow is as follows:

- A federal formula, based largely on income, allocates funds to the counties, through the states. The formula counts the number of children aged 5 to 17 from families below the Orshansky poverty level, children aged 5 to 17 from families receiving payments under AFDC of a specified amount, and children aged 5 to 17 supported by public funds. These counts come from census data, but are generally available only down to the county level.

- At the county level a state formula allocates funds to districts (where districts and counties are coterminous, this step is ignored). The state formula may use a different definition of poverty from that in the federal formula, but the definition must meet Title I statutes. In essence, the formula requires data that the state considers to best reflect the current and accurate distribution of eligible children from low-income families. In many states these data are very similar to those used in the federal formula, but several states use only AFDC counts, free-lunch counts, local survey, old census counts, or local knowledge of poverty concentrations.

- The districts, in turn, must determine the schools in which Title I services are to be provided and the level of services allocated to each school. The district must identify eligible schools, based for the most part on counts of poverty children in the attendance area, and then select the schools that will receive the services. The district can count low-income children, by attendance area, in any of the ways available to the states, or by school surveys, or health, housing or, employment statistics. The method employed may be a combination of any or all of these counting methods, but must be applied to all schools consistently. Having ranked the schools, each district must select those to receive the Title I services through another set of procedures that generally prohibit the 'skipping' of one school in need for another of lesser need, and that requires the Title I program to be of sufficient size, scope, and quality to give reasonable promise of success in meeting the goals of the program.

After target schools are identified from those eligible, the district must determine what kinds of students to serve (grades or educational problems), how many to serve of each kind, and the amount of funds to be allocated to each student's services. The only requirement is that children most in need of special assistance are to be served.

It is at the point where districts determine school eligibility that we focus our attention. Districts generally use 1970 census counts, current AFDC counts, and/or counts of free-lunch recipients—whichever are available—as the indicators of poverty concentration for schools. An NIE survey (NIE, 1977) collected counts of poor children in all schools in 100 Title I districts in 1975-76 according to three measures. 1970 census, enrolled AFDC children, and free-lunch recipients. The numbers of children identified by the three methods were correlated to see how similar they were. The correlation between census counts and AFDC children was +.58, between census and free lunch, +.51; and between AFDC and free lunch, +.81. The AFDC and free-lunch counts are the most similar, but the accuracy of any single method remained undetermined.

In the present study, we attempted to obtain accurate data on poverty concentrations of schools, and then to compare these data with the kinds of data typically used in practice, i.e., data in the form of principals' reports. As part of the Sustaining Effects Study (SES), principals were asked to estimate the percentages of poor students in their schools and teachers judged the poverty status of their students. We then examined how well these estimates related to accurate information derived from home interviews with parents. We also checked to see if certain characteristics of schools and principals influence these relationships.

THE ESTIMATES OF SCHOOL POVERTY

Information about the concentration of poor students was obtained for 242 nationally representative schools. The data were obtained from home interviews of parents and from judgments made by principals and teachers.

The Accurate Estimates. One of the mandates of the SES was to obtain accurate information on family economic status for a representative subsample of students. Between 15 and 200 home interviews were completed with parents of children participating in each school of the representative sample. Data from the 15,579 completed interviews could be projected to population estimates because of the methods used in constructing the sample. Income data from the interviews, fully described by Breglio, Hinckley, and Beal (1978) could also be transformed into standard indexes of poverty, and the percentage of poor students in each school could serve as an unbiased and accurate estimate of the school's poverty.

The index of poverty is referred to as 'current low income,' which is the percentage of students in each school meeting the criteria used by the federal government to allocate Title I funds to counties ('Federal Allocation Formula'). As noted above, these criteria are based on the number of families receiving AFDC, the number below the Orshansky poverty level, and the number of students residing in foster homes and institutions for the neglected or delinquent. Based on their national projections, Breglio et al. estimated 4.2 million such students, or 20.9 percent of the population of elementary school students.

Estimates by Principals. During the previous year (the 1975-76 school year), principals were asked to report poverty estimates for their schools. They responded to the following request:

> To estimate the precentage of students in your school who meet ESEA Title I poverty criteria, please answer (a) or (b), whichever is easier for you to estimate.
>
> (a) Pecentage of students in your school who meet the district's ESEA Title I poverty criteria ______%.
>
> (b) Percentage of students in your school who are eligible to receive free lunches ______%.

Thus, the principals estimated either the percentage of students meeting district Title I school poverty criteria ('Principal Report of Title I Eligible') or the percentage eligible for free lunch ('Principal Report of Free-Lunch Eligible'). Eighty-two percent of the principals chose to estimate the percentage eligible to receive free lunch, whereas 47 percent estimated the percent of Title I eligible. (Thirty-one percent of the principals reported both estimates.)

Estimates by Teachers. Each homeroom teacher in each school estimated the poverty status of each of his/her students. The teachers' estimates were made in response to the following item from the Student Background Checklist:

> To your knowledge, does the student participate in a free or reduced-priced lunch or breakfast program? (Use your own observation or school records. Do not ask the student.)
>
> • Yes • No

Responses were aggregated within schools to obtain estimates of school poverty that we call 'Teacher Report of Free-Lunch Participation.'

Estimates at the District Level. The criteria used to allocate Title I services to schools vary by district. Most districts employ multiple criteria, with the average district using

The three most frequently used criteria (Wang, Hoepfner, Zagorski, Hemenway, Brown, and Bear, 1978) are, in order of frequency:

1. Free or reduced-price lunch count,
2. AFDC enrollment, and
3. Census data on family income.

Because each district indicated which criteria were used to allocate Title I services to schools, and because we had independent information on each criterion, we could aggregate all relevant criteria to the school level and estimate for each school the percentage of students meeting the district's Title I school-eligibility criterion ('District Allocation Formula').

Among the various estimates of poverty, we would expect the 'Principal Report of Free-Lunch Eligible' to be highly related to the Teacher Report of Free-Lunch Participation,' since the percentage of students in each school eligible to receive free lunches, as estimated by the principal, should be comparable to the percentage participating in the lunch or breakfast programs as reported by the teachers, albeit a year later. The principal's estimate should also be highly related to the 'District Allocation Formula,' as most districts use free-lunch counts as a criterion. The 'Principal Report of Title I Eligible' should be related to the 'District Allocation Formula,' because, as noted above, principals were directed to make this specific poverty estimate. This principal's estimate should also be strongly related to the two poverty estimates based on free lunches because they share several common aspects in their definitions.

Additional analyses will determine whether specific school or principal characteristics influence the relationships between the various poverty estimates. Specifically, we consider the following characteristics that might be suspected to reduce accuracy:

- Stability of the student body
- Type of compensatory-education funding received
- Urbanism of school
- Minority concentration of school
- Percentage of students bused for racial balance
- Enrollment of school
- Principal's tenure in the school

RELATIONS AMONG THE POVERTY ESTIMATES

As may be seen in Table 2-1, principals' and teachers' estimates of students who are eligible for or participating in free-lunch programs (C and D, respectively) are most highly related to the poverty index we believe is most accurate; namely our data collected through home interviews and used in the Federal Allocation Formula (A). Principals' reports of students eligible for Title I (B) and the district formulas (E) are slightly less correlated with our accurate measure. Looking at it another way, the principals' estimates of free-lunch eligibility has very high concurrent validity with all the other estimates, and correlated highest with the most accurate measure, that obtained from the home interviews.

Principals' estimates of the percentages of students meeting district Title I eligibility criteria (B) are only moderately related to the percentages of students actually meeting these criteria (E) one year later (r = .54). These principal estimates were more highly related to the percentages of students meeting Federal Allocation Formula criteria (r = .69), and to the percentages of students participating in free or reduced-priced lunch or breakfast programs one year later (r = .78). Hence, the principal's Title I eligibility estimate is a much better indicator of free-lunch participation than it is of the percentage of students meeting the criteria established by the district for the school's Title I eligibility. This is not surprising, as the latter estimate is complex and generally includes several different criteria, with free-lunch participation as only one of its components. What is somewhat surprising is that the principal's free-lunch estimate (C) is a better predictor of the school's percentage of district Title I eligibles (E).

Not surprisingly, the free-lunch eligibility estimate is a better indicator of free-lunch participation. Hence, the principal's free-lunch estimate is superior to the principal's Title I estimate on several counts.

The means and standard deviations of the various poverty estimates, computed from all the cases available for each estimate, are presented in Table 2-2. We can see that home interviews (A) provide the highest average estimate of school poverty with district formulas being very similar. Principals' reports of Title I eligibles result in the lowest mean estimate. Comparing these means to the correlations in Table 2-1, we can see that, although principals' estimates of free-lunch eligibles are most highly correlated with the Federal Allocation Formula, they are considerably lower. (Correlation coefficients, of course, are insensitive to differences in the overall levels of the variables being correlated.) Of course, this consistent difference is not critical in cases where the estimate is applied uniformly across a set of schools, as when ranking them for Title I eligibility.

Table 2-1

Intercorrelations Among Poverty Estimates*

Poverty Estimate	A	B	C	D	E
A. Federal Allocation Formula**	113	194	238	238	
B. Principal Report of Title I Eligible	.69		73	113	113
C. Principal Report of Free-Lunch Eligible	.76	.79		194	194
D. Teacher Report of Free-Lunch Participation	.75	.78	.88		238
E. District Allocation Formula	.68	.54	.77	.74	

*Correlation coefficients below the diagonal; numbers of schools contributing to their calculation above. Sample sizes vary because principals chose to provide different estimates, and because some districts did not participate in Title I and had no allocation formula.

**This estimate, based on home interviews, is considered to be the most accurate.

Table 2-2

Means and Standard Deviations of Poverty Estimates

Poverty Estimate	N	Mean	S.D.
Federal Allocation Formula (A)	238	39 9	19.1
Principal Report of Title I Eligible (B)	113	28.8	23.8
Principal Report of Free-Lunch Eligible (C)	194	33.5	25.8
Teacher Report of Free-Lunch Participation (D)	238	38.5	28.7
District Allocation Formula (E)	238	39.8	25.9

Comparing selected means with a means with a *t*-test for paired observations, we find that the principal *underestimates* the percentage of Title I poverty eligibles, as compared to the percentage eligible by district criteria ($t_{(112)} = 4.65$, $p < .001$). The principal also *underestimates* the percentage of free-lunch eligibles as compared to the percentage receiving free or reduced-price lunch or breakfast (the number eligible should, of course, exceed the number participating), as reported by the teachers one year later ($t_{(193)} = 4.78$, $p < .001$). The generality of statistically significant differences among the mean estimates confirms intuitive knowledge that none of the poverty estimates is perfectly comparable to another poverty estimate.

THE EFFECTS OF SCHOOL AND PRINCIPAL CHARACTERISTICS ON THE ESTIMATES

The relationship between the principals' estimates of free-lunch eligibility (C) and the teachers' estimates of free-lunch participation (D), was examined through cross-tabulation, controlling for various school and principal characteristics as third variables. Similar analyses were performed on the principals' Title I eligibility estimates and the measures of Title I eligibility based on district criteria.

The results are reported in Table 2-3. *Gamma* is a measure of association between two ordinal-level variables, yielding values not unlike those of a product-moment correlation coefficient. Zero-order *gamma* measures the relationship between two variables, e.g., Principal Title I Eligible estimate of poverty and District Allocation Formula estimate, without controlling for any other variables. First-order partial *gamma* measures the relationship between the two variables controlling for one other variable, e.g., Student Body Stability. By comparing the zero-order *gamma* with the partial *gamma*, one can determine whether the third variable moderates the relationship between the other two. As can be seen in the table, in no instances were school or principal characteristics found to moderate the relationships between the poverty estimates. This finding presents indirect evidence that the characteristics do not influence the levels of the poverty estimates, either. However, if the characteristics influenced the levels of both poverty estimates similarly (which we don't consider very likely), no matter how strong the influence, the reduction in the partial *gamma* would be very small.

Table 2-3

The Effects of Controlling for School and Principal Characteristics on the Relationships Between Two Title I Estimates and Between Two Free-Lunch Estimates

Characteristic Controlled for	Principal Title I Eligible and District Allocation Formula	Principal Free-Lunch Eligible and Teacher Free-Lunch Participation
	First-Order Partial *Gamma*	First-Order Partial *Gamma*
(Zero-Order *Gamma*)	(.66)	(.91)
Student body stability	.62	.91
CE funding	.62	.91
Urbanism of school	.60	.91
Minority concentration	.58	.88
Busing for racial balance	.66	.90
School size (enrollment)	.65	.91
Principal's tenure	.66	.91

CONCLUSIONS

The estimates of school poverty are almost all highly related to one another, but they result in different levels of concentration of poverty. Using the Federal Allocation Formula applied to income data obtained from home interviews as the standard and accurate measure of school poverty, estimates by principals are significantly and consistently lower; estimates by teachers and by application of district allocation formulas are very close to the standard.

Principals' estimates of the percentages of students eligible for free-lunch programs are superior to their estimates of Title I eligibility for two reasons. First, the free-lunch estimate is a better indicator of the percentage of poor students in the school when poverty is defined by either district or federal allocation criteria. In addition, principals generally prefer to estimate the percentages of free-lunch eligibles.

CHAPTER 3: CRITERIA FOR DETERMINING TITLE I PARTICIPATION OF SCHOOLS

Ralph Hoepfner

Information on how districts determine which schools will participate in Title I are projected to the nation and contrasted with previous findings.

After federal Title I funds are allocated to counties on the basis of the number of children aged 5 to 17 from low-income families, those receiving welfare payments, and those supported by public funds, and the states have allocated the funds to school districts on the basis of some indication of the poverty level of the district's attendance area, the districts must select the schools to participate in the Title I program (see Moskowitz and Brown, this volume). Districts are to identify eligible schools on the basis of concentrations of poverty children in the schools' attendance areas. The methods used to obtain the poverty counts are not restrictive, so that districts can use a wide variety of more or less effective methods for ranking schools, so long as they provide current, accurate estimates of the numbers of eligible children from low-income families. Among those methods, districts can use census data, AFDC counts, free-lunch counts, free-breakfast counts, local surveys, old census counts, local knowledge about poverty concentrations, school surveys, or health, housing, or unemployment statistics.

Implications from the NIE Compensatory Education Study (NIE, 1977b) are that some of the more obscure methods are rarely used. Based on their nationally representative sample of 100 Title I districts, but not projected to population estimates, NIE reported that 67 percent used 1970 census data on family income; 66 percent used free-lunch and/or free-breakfast counts; 51 percent used counts of AFDC children; and 34 percent used other economic data, such as housing and unemployment statistics or local surveys. The numbers add up to more than 100 percent because many districts reported using more than one source of poverty data. (The Demonstration Study of NIE, 1977a, showed that, when given the opportunity, districts interested in trying something new usually abandoned the poverty-based procedures for selecting schools and adopted eligibility criteria based on achievement.)

The 219 districts in the Sustaining Effects Study (SES) provided information on the frequency of use of five methods of determining school eligibility for Title I that could be projected to national estimates. These data are presented in Table 3-1, along with the unweighted NIE estimates for purposes of comparison.

Table 3-1

District Methods for Determining Eligibility of Schools for Title I

Method	National Projections From 219 SES Districts		NIE Percentages (unweighted)
	Number	Percentage	
Census data on family income	5,934	44 4	67
AFDC enrollment	8,088	60 5	51
Free or reduced-price breakfast counts	1,714	12 8	66
Free or reduced-price lunch counts	10,450	78.2	
Number of neglected or delinquent children	4,084	30.6	—
Other economic data	—	—	34

The discrepancies between the weighted SES data and the unweighted NIE data (based on responses obtained in the same school year, but from differently defined samples: the SES sample includes all districts with grades 1-6; the NIE sample includes only districts participating in Title I) are considerable. Faulty assumptions about how districts determine school eligibility could result in inappropriate use of the NIE estimates. Our projected percentages, adjusted to consider only Title I districts, as NIE did, indicate much less reliance on census data (44.4 percent vs. 67 percent for NIE) and much more reliance on counts of free or reduced-price meals (the projected estimate is 78.2 percent for breakfast and/or lunch counts, compared to NIE's 66 percent).

Very few of the SES districts used only one method for determining school eligibility (the reason the percentages in Table 3-1 do not add to 100), so we also projected the incidence of combinations of methods. Table 3-2 presents the projected frequencies and percentages of all single methods and combinations of methods, based on the SES districts. (Because each respondent is counted in only one method combination, the percentages in Table 3-2 do add to the projected percentage of 93.5, the percentage of districts with grades in the 1-6 range that participate in Title I.)

Table 3-2

Methods and Combinations of Methods for Determining School Eligibility for Title I

Method: A = Census data on family income; B = AFDC enrollment; C = Free or reduced-price breakfast counts; D = Free or reduced-price lunch counts; E = Number of N or D children served

Method	Projected Numbers of Districts	Projected Percentages of Districts
D only	2,369	16.6
B and D	1,942	13.6
A, B, D, and E	1,490	10.4
A, B, and D	1,362	9.5
B only	1,194	8.4
D and E	658	4.6
A only	594	4.2
A and D	498	3.5
A and B	449	3.1
A, C, D, and E	389	2.7
A, B, and E	373	2.6
B, D, and E	369	2.6
A, B, C, D, and E	335	2.3
C and D	310	2.2
B, C, and D	281	2.0
A, D, and E	238	1.7
A, B, C, and D	159	1.1
E only	113	0.8
B, C, and E	113	0.8
C only	59	0.4
A,C, and D.	47	0.3
B and C	19	0.1
B and E	4	0.0
C, D, and E	3	0.0

Note Combinations that were not marked do not appear in this table, as their projected frequencies and percentages would be 0 We also project 926 districts, or 6.5 percent of the nation's districts with grades in the 1-6 range, do not participate in Title I .The percentages in the right column, therefore, add to 93.5 percent of the nation's total

CHAPTER 4. THE DISTRIBUTION OF TITLE I SERVICES

Ming-mei Wang, Ralph Hoepfner, Moraye B. Bear, Gerrie Smith

The results of a number of studies designed to inform Congress during the reauthorization hearings for Title I are reported. Title I was found to go to the poorest schools, with 'targeting' to schools differentiated by region and urbanism. Title I was also found to go to schools with higher percentages of low achievers. In the poorer schools, Title I services go more often to minority students, as would be expected from the average poverty and achievement status of minorities.

Before the reauthorization of ESEA Title I in 1978, the U.S. Department of Education requested information on the current status of the program. An evaluation of the current status would lead to recommendations for changes or additions in the law that would alter the program in the future. Much of the attention of the hearings on the reauthorization was expected to focus on the politically and educationally sensitive issue of funds allocation. Because considerations such as poverty, region, urbanism, minority status, and school size all play a role in decisions about allocations, the Department of Education wanted comprehensive national information about them. The various data bases of the SES were used for information needed in a manner more focused on policy needs than had been previously presented in the technical reports from the study. We therefore focused our analyses on the issues of poverty, achievement, minority status and school size.

SCHOOL POVERTY CONCENTRATION AND TITLE I

The Title I allocation formula prior to the 1978 reauthorization specified that a school's eligibility was to be based solely on its proportion of disadvantaged children. Districts rank their schools according to the proportions of children from families in poverty and then offer Title I programs in schools with the greatest concentrations of poverty. (It is only at the level of student selection that low achievement was to enter the selection process as a consideration.) Because the rankings were within districts, it was not necessarily the case that nationwide the poorest schools would participate in the program.

There are many reasons why this was so. For example, one district will have all schools at relatively low poverty concentrations; another district, with the same average poverty concentration might have some schools that are very poor and others that are not poor at all. This second district, in its ranking and selection procedures, might not be able to offer services to all of the very poor schools. We would expect, then, that some participating schools in the first district will have less poverty than some not participating in the second. Even if the distribution of poverty across schools were the same in two districts, one district might offer services in fewer schools with more intense programs while the other served more schools with less intense programs. Choices of the grades to be served or the subjects to be taught also reduce the exact correspondence between schools' participation in Title I and their poverty concentrations.

In the tables below, based on national projections from 4,750 respondents to the Principal Survey Questionnaire in 1975-76 (see Hoepfner, Zagorski, and Wellisch, 1977, for more information), we look at Title I and school poverty. From Table 4-1 we can see that the correspondence between concentration of poverty and being a Title I school is reasonably large. At the low level of poverty, half the schools offer Title I services but at the high level almost 85 percent do.

The superimposition of region onto the poverty distribution of Title I schools can serve to identify those regions in which the 'targeting' of schools is more or less closely aligned with the intentions of the Title I allocation formulas. Regions in which such targeting conforms well ought to have small percentages of low-poverty (0-19%) schools participating and much higher percentages of high-poverty (30-100%) schools. From Table 4-2 we can see

Table 4-1

Percentages of the Nation's Elementary Schools Offering Title I, by Poverty Concentration

Poverty Concentration	Title I Schools	Non-Title I Schools	Total
0-19% poverty enrollment	50 8	49 2	100.0
20-29% poverty enrollment	76 7	23.3	100 0
30-100% poverty enrollment	84 5	15 5	100.0
Total	67 5	32.5	100.0

Note Projected numbers supporting the percentages can be found in Table A of the Appendix.

Table 4-2

Projected Percentage of Schools Offering Title I, by Poverty Concentration and Geographic Region

Geographic Region	Poverty Concentration of Schools			Regional Totals
	0-19%	20-29%	30-100%	
New England	51 9	83.8	78 9	66.4
Metropolitan Northeast	61 3	86 8	83.8	71 3
Mid-Atlantic	55 9	71.2	89 4	71.0
Southeast	36 7	69 1	89.4	78.1
North Midwest	61 1	87 0	78.9	68.9
South Central	39 1	73 6	88.0	73.3
Central Midwest	63 5	93 4	81 6	73 8
North Central	42 2	76 8	75 2	55 4
Pacific Southwest	14.4	50 1	69 3	38.7
Pacific Northwest	49 5	73 7	95.8	65 0
Poverty Concentration Totals	50 8	76 7	84 4	67.5

Note: Projected numbers of schools can be found in Table A of the Appendix.

that no region displays any reversal in the order of percentages over the poverty categories. Half the regions have relatively more 'mid-level' poverty schools offering Title I than 'high' poverty schools. The regions appearing to conform most to allocation intentions are the Pacific Southwest, Southeast, and South Central; while those appearing to conform least are the North Midwest, Central Midwest, and Metropolitan Northeast.

Table 4-3 shows the correspondence between the proportion of poverty and the presence of Title I programs as it varies across seven categories of urbanism. Here one sees that the pattern across all degrees of urbanism is the same as the overall national picture, except for rural districts. In those districts there is a slight reversal in which a greater percentage of mid-level poverty schools offer Title I than high-poverty schools.

The three categories of cities with populations over 50,000 conform best to the allocation intentions.

Table 4-3

Projected Percentage of Schools Offering Title I, by Poverty Concentration and Urbanism

Urbanism	Poverty Concentration of Schools			Urbanism Totals
	0-19%	20-29%	30-100%	
City of over 500,000	14.3	25.5	74.2	56.2
City of 200,000-500,000	12.9	48.1	73.0	49.2
City of 50,000-200,000	18.0	53.8	78.8	48.5
Suburb of city	45.9	66.6	73.7	51.5
City under 50,000	57 2	79.6	91.2	72 7
Rural area near city	68.4	84.6	89.2	76.8
Rural area not near city	66.2	91.2	88.1	81.3
Poverty Concentration Totals	50.8	76.7	84.4	67.5

Note: Discrepancies within and between tables are due to rounding errors.

STUDENT ACHIEVEMENT, SCHOOL POVERTY, AND TITLE I

Achievement and CE. After schools are identified as eligible to offer Title I programs, they are to provide services (purchased with Title I funds) to the educationally most needy students Due to the sequential nature of the process for getting funds and services targeted to students in various sites, and due to the fact that the criterion changes from an economic to an educational orientation, not all of the lowest-achieving students in the nation will participate For example, such students may attend schools that do not offer Title I; they may not be the lowest of the low-achieving in their school; or they may be in a grade that is not participating at their school.

Although we may not expect an exact correspondence between a student's achievement and participation in Title I, if students are grouped by their participation status, we would expect Title I students to rank very low. This expectation is borne out in Table 4-4, that presents unweighted data from the SES representative sample from the fall of the first year (1976) of the study.

For both reading and math, Title I students always have the lowest mean achievement, with non-compensatory students (in schools offering compensatory services only from sources other than Title I) always having the highest. It may be noted that the compensatory-education (CE) students always have the three lowest achievement means, and the non-CE students always have the highest. Table 4-4 also presents some other information of interest; looking down the columns, we can see quite clearly the results of the difficulties of selecting low achievers at grade 1 (where achievement tests are unreliable and achievement is only partly academic). The standard deviations in Table A of the Appendix confirm this point The grade 1 difficulties seem to be not so great for selecting participants in math CE programs, however. While there is a slight trend for the means of the CE students to be lower at the higher grades (cross-sectionally), the means of the non-participants seem to remain fairly constant. It is this cross-sectional phenomenon that has been labelled 'the widening achievement gap.'

Teachers' Judgments of Need for CE. Table 4-5 presents information on CE participation according to teachers' judgments of students' needs for CE, an acceptable student-selection criterion according to Title I regulations. Our data are based on judgments made for each student, separately for reading and math, by the homeroom teacher. We can see from Table 4-5 that more Title I students are judged needy by their teachers than any other group of students, both for reading and math. The ranking of the groups by judged need is exactly the same as the rankings by achievement in Table 4-4.

We have converging evidence that, in general, the lowest-achieving students are participating in Title I. The evidence is that teachers' judgments are rather credible, although we cannot assume that they would remain so if achievement test scores were not also commonly used We can also see from Table 4-5 that there are few trends over grade, except that more other-CE students in other-CE schools are judged to need CE as grade increases.

In a more theoretical vein, we investigated the variables that explain (in a regression sense) whether a student is selected for Title I and whether a student is judged as needing CE by the classroom teacher. The practical implications of this work concerned the advisability of recommending teacher judgment in lieu of other (more objective) methods for selecting participants. If, for example, we found that race/ethnicity was important in teachers' judgments of need but not important in explaining actual participation (it wasn't found to be a useful predictor of either), regulations for selecting students to receive Title I could be altered to reduce bias that might have been indicated by the finding. Table 4-6 presents the average variances accounted for by each of the stable predictors for both Title I participation and judged need for CE.

Generally the overall predictability of Title I participation is slightly lower than the predictability of CE in general (see Wang, Hoepfner, Zagorski, Hemenway, Brown, and Bear, 1978). When teachers' judgments are substituted as the criterion to be predicted, and participation status is not allowed to enter as a predictor, predictability is considerably higher.

For reading judgments, it seems that the best predictors are current achievement, previous year's receipt of CE, and economic status; for math judgments, the best predictors are the same and in the same order. The relative differences in the percentages of variance accounted for in the prediction of the two criteria do not support an interpretation that selection and judgment by teachers are based on different considerations.

Table 4-4

Average Achievement of Students, by CE Status and Grade

Reading CE Status	Reading Percentiles					
	Gr. 1	Gr. 2	Gr. 3	Gr. 4	Gr. 5	Gr. 6
Students in Title I Schools						
Title I students	39	26	24	25	22	22
Other CE students	42	41	37	36	35	33
Non-CE students	49	52	53	52	52	53
Students in Schools With Only Other CE						
CE students	42	33	33	33	27	26
Non-CE students	54	64	62	61	60	59
Students in Schools With No CE						
All students	48	48	52	53	53	51
Math CE Status	**Math Percentiles**					
	Gr. 1	Gr. 2	Gr. 3	Gr. 4	Gr. 5	Gr. 6
Students in Title I Schools						
Title I students	34	29	24	29	26	27
Other CE students	45	42	36	39	40	38
Non-CE students	49	49	48	50	49	50
Students in Schools With Only Other CE						
CE students	41	40	38	36	29	29
Non-CE students	55	59	58	56	57	56
Students in Schools With No CE						
All students	48	48	50	50	52	52

Note Standard deviations, sample sizes, and means calculated to two decimal places can be found in Table B of the Appendix.

Table 4-5

Percentages of Students Judged by Teachers as Needing CE, by Their CE Status

Grade	In Title I Schools: Title I Students		In Title I Schools: Other-CE Students		In Title I Schools: Non-CE Students		In Schools with Only Other CE: Other-CE Students		In Schools with Only Other CE: Non-CE Students		In Schools With No CE: All Students	
	Needy	Not Needy	Needy	Not Needy	Needy	Not Needy	Needy	Not Needy	Needy	Not Needy	Needy	Not Needy
Reading												
1	80	20	46	54	26	74	61	39	16	84	23	77
2	88	12	59	41	24	76	73	27	12	88	30	70
3	87	13	52	48	21	79	70	30	14	86	28	72
4	86	14	51	49	22	78	72	28	15	85	25	75
5	87	13	60	40	22	78	78	22	13	87	18	82
6	87	13	58	42	17	83	77	23	16	84	21	79
Across Grades	86	14	54	46	22	78	72	28	14	86	24	76
Math												
1	76	24	41	59	22	78	52	48	14	86	17	83
2	79	21	51	49	21	79	51	49	10	90	21	79
3	82	18	43	57	21	79	53	47	13	87	25	75
4	75	25	47	53	22	78	64	36	16	84	19	81
5	84	16	55	45	23	77	71	29	13	87	18	82
6	80	20	44	56	20	80	68	32	13	87	18	82
Across Grades	79	21	46	54	22	78	61	39	13	87	20	80

Note: Sample sizes supporting the unweighted percentages can be found in Table C of the Appendix.

Achievement, Poverty, and CE. Funds are distributed to schools based on poverty because it is believed that those schools will also have large numbers of low achievers. Therefore, it is important to note the correspondence between achievement and poverty at the school level. Students scoring at or below the 35th percentile on normed reading and math achievement tests were defined as 'low achievers' and were aggregated to provide school percentages of low achievers. Mean percentages were then projected for Title I and non-Title I schools by the three poverty categories. The projected means are presented in Table 4-7. The percentage of low achievers varies directly with concentrations of poverty. Title I schools have higher percentages of low-achieving students, except at the middle category of poverty. This middle category is, quite likely, in the general range where many districts have to stop selecting schools for Title I, so the reversal of percentages is not totally unexpected.

Table 4-6

Percentages of Variance of CE Selection and Judgment of Need Accounted for, by Student Characteristics

Predictor	CE Selection		Teacher's Judgment of CE Need	
	Reading	Math	Reading	Math
Teacher's judgment of need for CE	26.3%	17.3%	—	—
Previous year's CE receipt	4.4%	6.2%	24.9%	16.4%
Pretest achievement (percentile score)	0.3%	0.3%	15.5%	12.3%
Other language spoken in the home	0.1%	0.1%	0.4%	1.5%
Economic status index	0.1%	0.2%	1.4%	2.3%
Total variance accounted	31.1%	24.1%	42.2%	32.5%

Note Discrepancies are due to rounding errors. More complete statistical data on the regression analyses can be found in Tables D and E of the Appendix.

Table 4-7

Percentages of Low-Achieving Students in Title I and Non-Title I Schools, by School Poverty Concentration

Title I Participation	Poverty Concentration of Schools			Across Poverty Levels
	0-19%	20-29%	30-100%	
Title I schools	16.2	23.7	40.1	28.6
Non-Title I schools	15.6	25.5	35.8	20.3
Across Title I status	15.9	24.1	39.4	25.9

In Tables 4-8 and 4-9, we introduce region and urbanism into the tabulations of percentages of low achievers in schools with different degrees of poverty concentrations. It appears that the large cities and the rural areas in the New England, Mid-Atlantic, South Central, Central Midwest, and Pacific Northwest regions contribute to the higher low-achievement percentages for non-Title I schools in the middle poverty category. Comparison of figures for Title I and non-Title I schools in the right-hand columns of both tables shows that no region or type of area has Title I schools with a lower mean concentration of low achievers than non-Title I schools. To the extent that this comparison indicates how well targeting of Title I to schools with low-achieving students has been, the Pacific Southwest region and the very large cities do the best job, while the North Midwest and Central Midwest regions and suburbs do the worst.

Table 4-8

Percentages of Low-Achieving Students in Title I and Non-Title I Schools, by Geographic Region and Poverty Concentration

Region Title I Participation	Poverty Concentration of Schools			Overall
	0-19%	20-29%	30-100%	
New England				
Title I schools	16.1	18.6	29.7	21.9
Non-Title I schools	13.3	19.9	25.2	16.3
Metropolitan Northeast				
Title I schools	13.5	26.9	40.7	25.1
Non-Title I Schools	11 2	12.3	29.9	14.3
Mid-Atlantic				
Title I schools	15.0	24 1	39.4	28.1
Non-Title I schools	16 4	32 1	33.8	21.2
Southeast				
Title I schools	29 2	31 9	43.0	40.6
Non-Title I schools	19.3	29 2	42.9	29 3
North Midwest				
Title I schools	14.6	22.0	36.4	21.3
Non-Title I schools	15 0	22.4	35.5	18.4
South Central				
Title I schools	17 2	23 8	39 5	33.8
Non-Title I schools	16 6	28 8	36.7	23.8
Central Midwest				
Title I schools	16.4	21.3	32.2	21.6
Non-Title I schools	15.5	23 2	29.5	18.1
North Central				
Title I schools	18.2	18 4	35.3	22.8
Non-Title schools	12.4	16.1	22.8	13.9
Pacific Southwest				
Title I schools	23 6	32.1	51 4	41.8
Non-Title I schools	17 5	27 2	41 7	22.8
Pacific Northwest				
Title I schools	18.8	17.8	39.4	24.8
Non-Title schools	19.4	19.9	20.0	19.5
National Totals				
Title I schools	16 2	23.7	40.1	28.6
Non-Title I schools	15.6	25.5	35.8	20.3

Note Projected numbers of schools can be found in Table A of the Appendix.

Table 4-9

Percentages of Low-Achieving Students in Title I and Non-Title I Schools, by Urbanism and Poverty Concentration

Urbanism Title I Participation	Poverty Concentration of Schools 0-19%	 20-29%	 30-100%	Overall
City of over 500,000				
Title I schools	17.3	29 6	57.1	53.7
Non-Title I schools	17.0	30.9	39 3	28.3
City of 200,000-500,000				
Title I schools	24 3	26 9	48.9	43.4
Non-Title I schools	19 8	23.8	35.3	24.7
City of 50,000-200,000				
Title I schools	19.6	30.6	40 5	35.6
Non-Title I schools	15.5	24.6	34.1	19.8
Suburb of city				
Title I schools	14.0	24.8	33.4	18.9
Non-Title I schools	13 9	24.7	45.1	16.8
City under 50,000				
Title I schools	14.8	21.8	35.0	24.7
Non-Title I schools	15.6	22 8	34.4	18.7
Rural area near city				
Title I schools	17 6	21 7	36.8	24.1
Non-Title I schools	16 1	30.1	35.5	20.2
Rural area not near city				
Title I schools	18.1	24 3	38.4	29 7
Non-Title I schools	16.4	26.0	31.7	21.8
National Totals				
Title I schools	16 2	23.7	40.1	28.6
Non-Title I schools	15.6	25.5	35.8	20.3

Note Projected numbers of schools can be found in Table A of the Appendix.

Tables 4-8 and 4-9 tell us something about how well Title I is targeted to schools in terms of the low achievement of their students (a criterion that does not directly enter into school selection), but tells us nothing about the students who actually participate in Title I. To get a picture of how well CE services are targeted to students, we altered our concern to the percentage of regular achievers (scoring above the 35th percentile) who participate in CE. We limited our search to the 213 CE schools in the first year of the study. The percentages of schools in each region or area of urbanism with more than 10 percent of their regular achievers in CE programs are presented in Tables 4-10 and 4-11. The data in these tables are unweighted, and based on small samples, but they do reveal some trends.

In almost all cases, there is a large percentage of Title I schools in the 'more-than-10-percent-of-regular-achievers-in-CE' category. Although we cannot rule out the possibility that Title I services are just badly mistargeted, we believe that the bad record for Title I is largely caused by the nature of the CE categories and the programs with which it is being compared. Our category of Title I schools includes schools that also have other CE, so the poor targeting may be worsened by those additional CE programs. Other-CE programs, when they are not layered on top of Title I (as they aren't, by definition of our catergories) tend to be much smaller than Title I programs, and therefore serve fewer students. We would expect that smaller programs are easier to target well, even if their selection criteria are not as clear as those of Title I. Our rationale is confirmed, for example, in Table 4-10, where the Pacific Southwest has the worst targeting for non-Title I schools. This region, of course, contains California, which has one of the most extensive state CE programs in the nation. Other things being equal, the larger the program in the school, the more we can expect to see mistargeting of CE services to regular achievers.

Table 4-10

Percentages of Title I and Other-CE Schools, by Region, With More Than 10 Percent of Their Regular Achievers Participating in CE Programs

Geographic Region	Reading		Math	
	Title I	Other-CE	Title I	Other-CE
New England	69	25	0	0
Metropolitan Northeast	72	0	50	0
Mid-Atlantic	76	18	47	17
Southeast	35	50	30	0
North Midwest	33	29	33	14
South Central	44	0	37	0
Central Midwest	47	20	41	20
North Central	47	20	13	20
Pacific Southwest	64	63	73	63
Pacific Northwest	58	0	25	0

Note Sample sizes supporting the unweighted percentages can be found in Table F of the Appendix

Table 4-11

Percentages of Title I and Other-CE Schools, by Urbanism, With More Than 10 Percent of Their Regular Achievers Participating in CE Programs

Urbanism	Reading		Math	
	Title I	Other-CE	Title I	Other-CE
City of over 500,000	80	0	80	0
City of 200,000-500,000	50	40	50	40
City of 50,000-200,000	50	8	21	0
Suburb of city	15	14	23	14
City under 50,000	56	33	24	8
Rural area near city	47	29	35	29
Rural area not near city	59	67	41	67

Note Sample sizes supporting the unweighted percentages can be found in Table F of the Appendix

Another way to gain an understanding of the nature of schools with different success in targeting CE to low-achieving students is to calculate averages of certain school characteristics for each category of targeting success and then to compare the averages to see if there are distinct differences among the categories. Eight school characteristics are so treated in Table 4-12. Keeping in mind that the leftmost column signifies 'good' targeting of CE (few regular-achieving students participate) and the rightmost column signifies bad targeting, we can see that:

- schools in larger districts tend to serve more regular achievers;
- poorer schools and schools in poorer districts tend to serve more regular achievers;
- schools with higher rates of parent/community involvement tend to serve more regular achievers: and
- schools with more minority students tend to serve more regular achievers.

Differences according to schools size and expenditures are not strong, except that schools in the highest category of poor targeting have very low per-pupil CE expenditures (probably caused by a greater number of participants overall, who serve as the divisor in creating the per-pupil index).

MINORITIES, SCHOOL POVERTY, AND TITLE I

So far in our consideration of the distribution of Title I, we have addressed only one educational concern—achievement. Issues of poverty, region, and urbanism are political. The minority status of participating students is another political concern. Even though Title I is not, *prima facie*, a minority program (and such considerations do not appear directly in the laws or regulations), to the extent that minorities reside in areas of high poverty, and to the extent that they are low achievers, we would expect their participation rates to be higher than for non-minority students. The students we classified as being minority include

Table 4-12

School Characteristics for Title I Schools With Various Percentages of Regular-Achieving Students Participating in CE Reading Programs

School Characteristics	Percentage of Regular Achievers Participating in CE Reading Programs			
	0-10%	11-20%	21-40%	41-100%
Number of elementary schools in district	13.3	12.3	45.4	49.3
District's percentage of poverty students	24 4	26.4	32 1	31.9
School enrollment in grades 1-6	306	313	294	329
Parent/community involvement (index)	39 3	41.2	41.8	48.1
School's percentage of minority students	19.4	19.3	40.2	49.2
School's percentage of poverty students	10.9	14.7	23.7	24.3
Current per-pupil expenditures	$1,170	$1,274	$1,179	$1,153
Per-pupil CE expenditures	$ 353	$ 359	$ 362	$ 248

Note Standard deviations, sample sizes, means calculated to two decimal places, and descriptions of the expenditure data can be found in Table G of the Appendix.

American Indian, Asian American, black, Hispanic, and all other non-whites. The percentage of minority students in each school was calculated as the sum of the principal's estimates of the percentage of students in each component racial/ethnic group.

From Table 4-13 we can see clear evidence from the rightmost column that minorities do tend to attend schools in higher poverty areas. In the lowest poverty area, Title I schools tend to have lower concentrations of minority students than non-Title I schools, but that is reversed in the higher poverty areas where Title I schools have higher concentrations of minorities. (We could conjecture that minority students in low-poverty areas are children in upwardly mobile families who excell their non-minority neighbors in motivation and achievement.) When these percentages are further analyzed by region, as in Table 4-14, we can see the almost universal trend that low-poverty Title I schools have lower concentrations of minority students than their non-Title I counterparts, but the reverse is true at the

Table 4-13

Projected Mean Percentages of Minority Students in Title I and Non-Title I Schools, by Poverty Concentration

Poverty Concentration	Title I Schools	Non-Title I Schools	Overall
0-19% poverty enrollment	5.3	8.8	7.0
20-29% poverty enrollment	10 3	20.3	12.7
30-100% poverty enrollment	40 5	35.8	39.8
Overall Poverty Levels	22 4	14.9	20.0

Note Projected numbers supporting the percentages can be found in Table A of the Appendix.

73

Table 4-14

Percentages of Minority Students in Title I and Non-Title I Schools, by Geographic Region and Poverty Concentration of Schools

Region Title I Participation	Poverty Concentration of Schools 0-19%	20-29%	30-100%	Overall Poverty
New England				
Title I schools	2.3	1.7	22.8	10.1
Non-Title schools	3.0	5.6	19.1	6.5
Metropolitan Northeast				
Title I schools	5.6	6.5	55.2	22.8
Non-Title I schools	5.4	33.3	47.4	14.0
Mid-Atlantic				
Title I Schools	4.0	9.9	36.5	20.5
Non-Title I schools	9.5	27.7	30.9	15.3
Southeast				
Title I schools	16.6	14.8	38.2	33.7
Non-Title I schools	14.3	17.7	35.4	22.2
North Midwest				
Title I schools	3.5	6.9	37.0	12.3
Non-Title I schools	5.9	15.8	37.3	11.0
South Central				
Title I schools	12.9	19.6	50.6	40.2
Non-Title I schools	14.3	25.5	47.5	24.8
Central Midwest				
Title I schools	3.4	5.9	26.3	9.7
Non-Title I schools	3.5	1.4	16.3	5.4
North Central				
Title I schools	5.3	9.4	32.6	13.7
Non-Title I schools	4.3	13.9	25.4	7.6
Pacific Southwest				
Title I schools	18.4	25.4	53.9	40.7
Non-Title I schools	16.7	26.9	41.1	22.1
Pacific Northwest				
Title I schools	5.3	7.6	28.0	12.8
Non-Title I schools	5.9	6.4	0.0	5.8
National Totals				
Title I schools	5.3	10.3	40.5	22.4
Non-Title I schools	8.8	20.3	35.8	14.9

Note Projected numbers of schools can be found in Table A of the Appendix high-poverty level. The major exception occurs in the North Midwest region, where the minority concentrations for Title I schools are always lower than for non-Title I schools. If the minority percentage for non-Title I schools is subtracted from that for Title I schools (rightmost column of Table 4-14), although the size of the difference depends in part on the overall regional minority concentration, we can still see that schools with high concentrations of minorities are most served by Title I in the South Central and Pacific Southwest regions, relative to non-Title I, and least served in the North Midwest.

Table 4-15 presents average concentrations of minority enrollments over the seven categories of urbanism It is clearly in the cities, at all levels of poverty, that Title I schools

Table 4-15

Percentages of Minority Students in Title I and Non-Title I Schools, by Urbanism and Poverty Concentration of Schools

Urbanism Title I Participation	Poverty Concentration of Schools			Overall Poverty
	0-19%	20-29%	30-100%	
City of over 500,000				
Title I schools	33.2	43.9	85.0	80.3
Non-Title I schools	21.6	39.8	57.4	39.1
City of 200,000-500,000				
Title I schools	22.8	29.1	67.7	57.9
Non-Title I schools	14.2	18.4	44.2	23.0
City of 50,000-200,000				
Title I schools	13.7	25.8	50.2	40.7
Non-Title I schools	10.2	22.7	40.6	17.0
Suburb of city				
Title I schools	5.6	18.1	39.6	13.1
Non-Title I schools	8.4	23.1	36.6	11.3
City under 50 000				
Title I schools	4.4	9.0	33.6	17.6
Non-Title I schools	7.8	14.9	25.9	10.7
Rural area near city				
Title I schools	3.7	9.5	28.1	12.1
Non-Title I schools	3.[illegible]	10.7	27.7	7.5
Rural area not near city				
Title I schools	4.4	4.9	26.0	15.3
Non-Title I schools	5.1	10.8	16.3	8.9
National Totals				
Title I schools	5.3	10.3	40.5	22.4
Non-Title I schools	8.8	20.3	35.8	14.9

Note Projected numbers of schools can be found in Table A of the Appendix have high minority enrollments, absolutely with regard to overall population expectations, and relatively with regard to non-Title I schools. At the highest poverty level, regardless of urbanism, Title I schools enroll higher percentages of minorities than non-Title I schools, but the same is not true at the lower poverty levels.

SCHOOL SIZE

The last characteristic we look at with respect to schools that offer Title I is their size. One long-standing requirement incorporated into Title I law and regulations is termed 'comparability.' It requires that services provided by state and local funds be comparable throughout the participating district (with certain exceptions). Comparability makes it impossible for districts to juggle funds and services among schools so as to effect a reduction in state or local support to schools that receive Title I funds resulting, for example, in a lowering of the property tax rate at the expense of the Title I grant. One of the problems with the idea of comparability is that provision of the same services may involve different costs for small or large schools, or for schools in near or remote locations. 'Exceptional schools,' as we will call them, result in complications for districts, that, for example have a few small and remote schools. To maintain exact comparability the district would have to make costs equal in all similar schools, either to the higher costs of the exceptional schools, thus raising district tax requirements, or to the lower costs of the large and near schools, thus making it impossible for the small and remote schools to provide adequate educational services.

In the example above, we discussed school size and remoteness as contributors to district problems in meeting comparability requirements. Because school size is easier to assess than remoteness, the Congress wanted information on the distribution of small (less than 100 students) and large schools, so it would have an idea of the results of writing an exception into the law whereby small schools could be excluded from district comparability standards. In Table 4-16 we tabulate the percentage of each region's schools that are small, and in Table 4-17 we do the same for the categories of urbanism. We also report the percentages of students enrolled in the small schools.

From Table 4-16 we can see that the North Central region has the highest percentage of small schools, so the districts in that region could expect to see some simplification of their comparability reports. The Pacific Northwest, Central Midwest, and New England regions also have large percentages of small schools. As expected, Table 4-17 shows rural areas as having higher percentages of small schools.

When we look at the distribution of large schools (enrollments of 100 or more) over region or urbanism and poverty, as in Tables 4-18 and 4-19, we can see that comparability requirements (expected to fall most heavily on districts with large and poor schools) will be greatest for the Southeast and South Central regions and for the large cities where there is simply greater poverty, and will be least for the North Central and Pacific Northwest regions and for suburbs and rural areas near cities.

Table 4-16

Percentages of the Nation's Elementary Schools That Are Small (Enrollments in Grades 1-6 Less Than 100) and of Their Enrollments, by Geographic Region

Region	Percentage of Schools	Percentage of Enrollment
New England	16.1	3.9
Metropolitan Northeast	0.6	0.1
Mid-Atlantic	5.6	1.2
Southeast	3.5	0.7
North Midwest	5.0	1.1
South Central	11.7	2.3
Central Midwest	16.1	4.2
North Central	36.3	7.2
Pacific Southwest	7.2	1.2
Pacific Northwest	19.2	3.6
Total	9.3	1.7

Note Projected numbers on which the percentages are based can be found in Table H of the Appendix.

Table 4-17

Percentages of the Nation's Elementary Schools That Are Small (Enrollments in Grades 1-6 Less Than 100) and of Their Enrollments, by Urbanism

Urbanism	Percentage of Schools	Percentage of Enrollment
City of over 500,000	0.8	0.1
City of 200,000-500,000	2.5	0.5
City of 50,000-200,000	1.5	0.3
Suburb of city	1.6	0.3
City under 50,000	3.7	0.8
Rural area near city	10.3	2.3
Rural area not near city	26.5	6.7
Total	9.3	1.7

Note Projected numbers on which the percentages are based can be found in Table I of the Appendix

Table 4-18

Percentages of Each Region's Schools That Are Large (Enrollments of 100 or More in Grades 1-6), by Poverty

Region	Poverty (Free Lunch) Concentration				
	0-25%	26-35%	36-50%	51-75%	76-100%
New England	58.9	8.1	7.7	5.6	3.6
Metropolitan Northeast	69.2	7.7	7.7	7.9	6.8
Mid-Atlantic	54.7	14.3	11.6	10.1	3.8
Southeast	25.8	18.1	21.0	19.3	12.5
North Midwest	73 3	8.7	6.5	3.1	3.4
South Central	34.0	13 0	17.5	14.2	9.7
Central Midwest	63.2	9 1	5.9	3.2	2.5
North Central	51 6	6.4	2.2	1.7	1.8
Pacific Southwest	58.8	10.2	8.1	10.2	5.4
Pacific Northwest	65.4	8.0	4.7	3.8	2.2
Total	55.1	11 0	10.4	8.5	5.7

Note Projected numbers on which the percentages are based can be found in Table J of the Appendix.

Table 4-19

Percentages of Schools in Each Category of Urbanism That Are Large (Enrollments of 100 or More in Grades 1-6), by Poverty

Urbanism	Poverty (Free Lunch) Concentration				
	0-25%	26-35%	36-50%	51-75%	76-100%
City of over 500,000	28.6	9.7	13.0	18.2	29.7
City of 200,000-500,000	45.5	11.4	9.6	19.1	11.8
City of 50,000-200,000	55.0	12.5	14.1	11.3	5.8
Surburb of city	85 4	6.0	3.3	3.1	0 6
City under 50,000	60.7	13.1	12.0	7.3	3.2
Rural area near city	63.8	10.4	8.8	5.4	1.2
Rural area not near city	36.3	11.6	11.2	8.8	5.6
Total	55.1	11.0	10.4	8.5	5.7

Note Projected numbers of which the percentages are based can be found in Table K of the Appendix.

APPENDIX

SUPPLEMENTAL TABLES FOR CHAPTER 4

Table A

Projected Numbers of Schools Participating in Title I, by Poverty Concentration, Geographic Region, and Urbanism

Geographic Region / Urbanism	Poverty Concentration: 0-19% Poverty		20-29% Poverty		30-100% Poverty		Total	
	Title I	Non-Title I	Title I	Non-Title I	Title I	Non-Title I	Title I	Non-Title I
New England	1,144	1,058	665	128	1,144	306	2,953	1,492
Metropolitan Northeast	1,827	1,155	610	93	1,271	246	3,708	1,494
Mid-Atlantic	1,716	1,353	739	299	2,238	265	4,693	1,917
Southeast	538	927	853	381	5,727	682	7,118	1,990
North Midwest	5,322	3,383	1,820	271	2,290	611	9,432	4,265
South Central	683	1,062	879	316	3,523	479	5,085	1,857
Central Midwest	1,610	924	864	61	813	184	3,287	1,169
North Central	834	1,145	489	148	481	159	1,804	1,452
Pacific Southwest	442	2,621	552	549	1,392	617	2,386	3,787
Pacific Northwest	716	731	477	171	524	23	1,717	925
Total	14,833	14,360	7,949	2,417	19,403	3,571	42,185	20,348
City of over 500,000	143	855	115	336	2,298	799	2,556	1,990
City of 200,000-500,000	104	698	184	199	908	336	1,196	1,233
City of 50,000-200,000	532	2,430	459	395	2,235	601	3,226	3,426
Suburb of city	3,264	3,849	711	357	793	283	4,768	4,489
City under 50,000	4,787	3,586	2,592	663	5,341	517	12,720	4,766
Rural area near city	2,503	1,154	1,078	196	1,494	182	5,075	1,532
Rural area not near city	3,501	1,789	2,810	271	6,333	854	12,644	2,914
Total	14,833	14,360	7,949	2,417	19,403	3,571	42,185	20,348

N... ected numbers, subtotals, and totals are subject to small rounding errors.

Table B

Mean Reading and Math Percentile Scores, by CE Participation and Grade*

Student CE Participation Status		Reading						Math					
		Gr. 1	Gr. 2	Gr. 3	Gr. 4	Gr. 5	Gr. 6	Gr. 1	Gr. 2	Gr. 3	Gr. 4	Gr. 5	Gr. 6
Title I student in Title I school	Mean	39.32	26.02	24.33	25.10	21 97	22.40	34.25	29 35	24.37	29.21	25 92	27.49
	S D	24.97	18.74	17.86	18 14	17.80	17.37	23 53	23.05	20.84	21.97	20.55	21.47
	N	1,602	1,762	1,880	1,540	1,423	1,163	850	858	1,006	895	819	657
Other-CE student in Title I school	Mean	41.99	40.57	36.84	35.67	34.60	32.76	45.17	42.42	35.62	39.48	39.78	38.31
	S D	25 76	27.63	27.25	26.11	25.46	25.42	25.80	28.61	26.51	27.16	28.21	27.83
	N	815	826	706	682	640	724	731	626	532	626	539	634
Other-CE student in Other-CE school	Mean	41 88	32.88	33.25	32.55	27.04	26 22	41.27	39.97	38.44	35.65	28.98	28.79
	S.D.	30.17	23 66	24.58	23.88	20 61	20 82	28.67	29.64	27.75	26.49	24.71	22.38
	N	533	588	558	527	542	541	361	306	389	362	404	424
Non-CE student in Title I school	Mean	49.24	52.29	52.71	51 73	51.84	53 31	48.73	48 61	48.25	49.72	48.94	49.50
	S D	28.85	28 44	27.60	27.81	27 72	27.87	28.69	28.45	28.01	28.25	27.92	28.44
	N	6,613	5,462	5,278	5,434	5,622	5,962	7,411	6,545	6,322	6,124	6,326	6,546
Non-CE student in Other-CE school	Mean	54.41	64.18	61.77	60.92	60 05	58.54	54.65	58.91	58.36	56.28	56.79	55.94
	S.D.	29 94	26.17	26.87	27 50	26.68	27.50	29.38	27.38	28.16	29.29	28.35	28.80
	N	2,786	2,441	2,489	2,487	2,640	4,334	2,957	2,716	2,657	2,646	2,773	4,440
Non-CE student in Non-CE school	Mean	48.26	48 17	51.75	52.50	53.36	51 40	47.83	47.59	49.64	50.22	52.11	52.20
	S D.	30.37	28.13	28.02	28.94	29.62	28.80	29.85	28 32	28.00	28.02	29.71	29.06
	N	1,268	1,111	1,160	1,184	1,250	1,662	1,264	1,104	1,158	1,183	1,247	1,659

*The reading and math scores were obtained at the beginning of the school year (Fall, 1976).

Table C

Numbers of Students Judged by Teachers as Needing CE for Reading and Math, by CE Participation Status and Grade

Grade	In Title I Schools						In Schools with Only Other CE				In Schools With No CE	
	Title I Students		Other-CE Students		Non-CE Students		Other-CE Students		Non-CE Students		All Students	
	Needy	Not Needy	Needy	Not Needy	Needy	Not Needy	Needy	Not Needy	Needy	Not Needy	Needy	Not Needy
Reading												
1	1,363	35[illegible]	393	464	1,751	5,089	337	215	448	2,451	306	996
2	1,628	217	513	353	1,337	4,317	444	162	300	2,203	349	802
3	1,670	249	387	357	1,143	4,324	397	171	354	2,262	336	881
4	1,344	227	361	346	1,228	4,352	379	150	398	2,166	300	912
5	1,297	187	401	267	1,276	4,501	427	120	336	2,359	236	1,051
6	1,030	152	426	313	1,020	5,093	426	129	707	3,699	345	1,337
Total	8,332	1,382	2,481	2,100	7,7[illegible]5	27,676	2,410	947	2,543	15,140	1,872	5,979
Math												
1	692	215	315	454	1,702	6,032	186	169	417	2,679	217	1,085
2	704	187	338	320	1,428	5,388	155	150	273	2,531	243	908
3	834	179	241	326	1,394	5,156	208	184	350	2,442	299	918
4	681	229	303	347	1,381	4,917	233	129	426	2,305	233	979
5	717	139	306	254	1,521	4,992	290	120	364	2,468	231	1,056
6	538	138	282	364	1,359	5,353	[illegible]93	138	597	3,933	305	1,377
Total	4,166	1,087	1,785	2,065	8,785	31,838	1,365	890	2,427	16,358	1,528	6,323

Table D

Correlations, Standardized Regression Coefficients, and Percentages of Variances Reduced by Each Successive Entry of Predictors of Student Selection for Title I Services for Reading and Math, by Grade

Predictor	Grade 1			Grade 2			Grade 3			Grade 4			Grade 5			Grade 6		
	r*	Beta	%V**	r*	Beta	%V**	r*	Beta	%V**	r*	Beta	%V**	r*	Beta	%V**	r*	Beta	%V**
Reading																		
Teacher judgment of need for reading CE	42	36	17 8	52	40	27 3	56	[illegible]	31 5	52	36	27 3	52	31	26 9	52	32	27 0
Previous year's reading CE receipt	22	10	1 0	40	21	3 8	47	23	4 1	46	24	4 2	50	29	6 0	51	31	7,1
Reading achievement (percentile score)	- 14	- 04	0 2	- 37	- 09	0 3	- 41	- 08	0 5	- 37	- 06	0 2	- 39	- 08	0 3	- 36	- 04	0 2
Other language spoken in the home	07	- 03	0 0	04	- 06	0 3	09	- 01	0 0	08	- 03	0 0	11	00	0 0	08	- 01	0 0
Economic status index	- 23	- 07	0 4	- 21	01	0 0	- 24	- 01	0 0	- 23	- 02	0 0	- 22	02	0 0	- 24	- 02	0 0
Multiple correlation	44			56			60			56			58			59		
Sample size		8 166			7 181			7 084			6 847			6 937			7 327	
Total % variance			19 45			31 77			36 04			31 75			33 27			34.27
Math																		
Teacher judgment of need for math CE	37	28	13 7	42	32	17 2	45	29	20 3	40	25	16 4	44	28	19 5	41	23	16 7
Previous year's math CE receipt	26	15	2 3	34	21	4 1	43	29	7 5	39	26	5 7	45	31	7 8	46	35	10 0
Math achievement (percentile score)	- 16	04	0 2	- 23	- 05	0 2	- 29	- 05	0 3	- 25	- 07	0 5	- 29	- 03	0 1	- 23	- 05	0,2
Other language spoken in the home	08	- 03	0 0	05	- 05	0 1	10	01	0 1	05	- 04	0 0	06	- 04	0 1	07	- 01	0 0
Economic status index	- 24	10	0 7	- 18	- 02	0 0	- 22	- 04	0 1	- 20	- 05	0 2	- 20	- 03	0 1	- 18	- 01	0 0
Multiple correlations	41			47			53			48			53			52		
Sample size		6 757			5 797			5,731			5 403			5,368			5,862	
Total % variance			16 98			21 67			28 34			22 80			27 60			26 93

*r = simple correlation of the predictor with Title I participation status of 1976-77 school year based only on students in schools participating in Title I (all non participating students were considered as not receiving CE)

**%V = percentage of additional variances reduced by entry of the predictor order of the predictors determined from original predictor orders with with CE participation as the criterion and based on the entire SES sample

Table E

Correlations, Standardized Regression Coefficients, and Percentages of Variances Reduced by Each Successive Entry of Predictors of Teacher's Judgment of CE Need for Reading and Math, by Grade

Predictor	Grade 1			Grade 2			Grade 3			Grade 4			Grade 5			Grade 6		
	r*	Beta	%V**	r*	Beta	%V**	r*	Beta	%V**	r*	Beta	%V**	r*	Beta	%V**	r*	Beta	%V**
Reading																		
Previous year's reading CE receipt	30	24	9 1	42	25	17 9	53	34	2- 9	56	37	30 9	58	37	33 3	55	38	30 3
Reading achievement (percentile score)	23	- 16	-4 9	- 61	50	24 9	- 61	- 44	19 3	- 58	- 39	14 3	- 62	- 41	16 0	- 56	- 37	13 5
Other language spoken in the home	17	05	1 3	17	03	0 2	17	05	0 3	15	- 02	0 0	19	.04	0 3	14	01	0 0
Economic status index	38	29	7 2	- 36	04	0 1	36	- 05	0 2	- 36	- 08	0 4	- 39	- 07	0 3	- 35	- 07	0 4
Multiple correlation	47			66			69			68			71			67		
Sample size		8 166			7 161			7 084			6 847			6 937			7,327	
Total % variance			22 56			43 09			[illegible]			45 61			49 90			44 30
Math																		
Previous year's math CE receipt	30	23	9 1	37	26	13 8	41	25	16 5	44	33	19 2	46	30	21 1	.43	33	18 8
Math achievement (percentile score)	33	21	5 5	43	32	12 3	50	- 37	14 9	- 45	- 32	11 7	- 55	- 40	17 1	- 43	- 29	10 5
Other language spoken in the home	19	05	2 0	18	05	1 8	16	05	2 1	12	- 01	0 6	18	05	1 5	13	00	0 8
Economic status index	39	27	6 1	- 34	- 15	1 8	33	- 11	0 9	- 32	- 15	1 7	- 36	- 12	1 1	- 33	- 16	1 9
Multiple correlations	50			54			59			58			64			57		
Sample size		6 757			5 797			5 731			5 403			5 368			5,862	
Total % variance			24 69			29 68			34 41			33 16			40 77			32 03

*r = simple correlation of the predictor with Title I participation status of 1976-77 school year based only on students in schools participating in Title I (all non participating students were considered as not receiving CE)

**%V = percentage of additional variance reduced by entry of the predictor order of the predictors determined from original predictor orders with CE participation as the and based on the entire SES sample

Table F

Numbers of Title I and Other-CE Schools, by Region and Urbanism, With More Than 10 Percent of Their Regular Achievers Participating in CE Programs

Region / Urbanism	Reading Title I	Reading Non-Title I	Math Title I	Math Non-Title I
New England	9	2	0	0
Metropolitan Northeast	13	0	9	0
Mid-Atlantic	13	1	8	1
Southeast	7	1	6	0
North Midwest	6	[illegible]	6	1
South Central	7	0	6	0
Central Midwest	8	1	7	1
North Central	7	1	2	1
Pacific Southwest	7	5	8	5
Pacific Northwest	7	0	3	0
Total	84	13	55	9
City of over 500 000	8	0	8	0
City of 200 000-500 000	3	2	3	2
City of 50 000-200 000	7	1	3	0
Suburb of city	2	2	3	2
City under 50 000	26	4	11	1
Rural area near city	8	2	6	2
Rural area not near city	30	2	21	2
Total	84	13	55	9

Note: Sample is based on 213 schools in the first year of the study that participated in any CE program. It can be seen, then, that 97 of the schools had more than 10 percent of their regular-achieving students participating in CE reading programs and 64 had more than 10 percent participating in their CE math programs. The remaining schools had fewer than 10 percent of their regular-achieving students in the respective CE program.

Table G

School Characteristics for Title I Schools With Various Percentages of Regular-Achieving Students Receiving CE Reading Services

School Characteristics		Percentage of Regular Achievers Receiving CE Reading Services: 0-10%	11-20%	21-40%	41-100%	Total
Number of elementary schools in district	Mean	13.34	12 30	45 40	49.25	21.82
	S D	19 27	22.56	107.66	107.54	58.55
	N	73	43	25	16	157
District's percentage of poverty students	Mean	24 40	26.44	32 12	31.94	26.96
	S D	20 90	16.17	21 18	14 76	19.28
	N	73	43	25	16	157
School enrollment in grades 1-6	Mean	305 67	312.56	294 28	328.69	308.09
	S D	161 54	236.67	150 31	235.01	189.72
	N	73	43	25	16	157
Parent/ community involvement	Mean	39 33	41.16	41 76	48 12	41.11
	S D	12 35	16.31	12 53	18 91	14 39
	N	73	43	25	16	157
School's percentage of minority students	Mean	19 41	19 30	40.16	49.19	25.72
	S D	27 77	31.57	39.66	39.48	33.71
	N	73	43	25	16	157
School's percentage of poverty students	Mean	10.92	14 67	23.72	24.25	15.34
	S.D.	10.35	16 31	23.35	26.66	17.38
	N	73	43	25	16	157
Current per-pupil expenditures	Mean	$1,170	$1,274	$1,179	$1,153	$1,198
	S D	$525	$476	$400	$386	$479
	N	73	43	25	16	157
Per-participant CE Expenditure	Mean	$353	$359	$362	$248	$345
	S D	$282	$219	$206	$163	$245
	N	73	43	24	16	156

Note Per participant expenditures were obtained by dividing the reported total CE expenditures by the reported number of participants Valid expenditure data were not available for one Title I school. The index of parent/community involvement has an overall mean of 42 1 and a standard deviation of 14.6.

86

Table H

Projected Distribution of the Nation's Elementary Schools With Grades in the 1-6 Range and the Grade 1-6 Enrollments, by Geographic Region and School Size

Region	Small Schools (Less than 100)		Large Schools (100 or more)		Total Population	
	Number	Enrollment	Number	Enrollment	Number	Enrollment
New England	714	44,150	3,731	1,096,448	4,445	1,140,594
Metropolitan Northeast	32	3,009	5,170	2,252,999	5,202	2,256,008
Mid-Atlantic	373	27,936	6,236	2,234,407	6,610	2,262,333
Southeast	315	23,213	8,792	3,507,054	9,108	3,530,259
North Midwest	690	46,944	13,007	4,251,762	13,698	4,298,699
South Central	812	52,791	6,131	2,206,679	6,943	2,259,462
Central Midwest	719	44,472	3,738	1,017,828	4,457	1,062,297
North Central	1,180	47,184	2,075	606,228	3,255	653,412
Pacific Southwest	446	26,046	5,726	2,238,137	6,173	2,264,178
Pacific Northwest	508	25,488	2,135	678,888	2,643	704,373
Total	5,790	341,232	56,741	20,091,988	62,534	20,431,590

Note Discrepancies between totals and actual sums are due to rounding errors

Table I

Projected Distribution of the Nation's Elementary Schools With Grades in the 1-6 Range and the Grade 1-6 Enrollments, by Urbanism and School Size

Urbanism	Small Schools (Less than 100)		Large Schools (100 or more)		Total Population	
	Number	Enrollment	Number	Enrollment	Number	Enrollment
City of over 500 000	36	3,306	4,511	2,308,943	4,547	2,312,248
City of 200 000-500,000	61	4,761	2,368	953,819	2,429	958,579
City of 50,000-200,000	97	7,314	6,555	2,377,694	6,651	2,385,007
Suburb of city	150	12,726	9,105	3,681,238	9,254	3,693,961
City under 50,000	652	43,433	16,833	5,692,448	17,485	5,735,872
Rural area near city	679	44,338	5,927	1,925,640	6,607	1,969,968
Rural area not near city	4,116	225,350	11,442	3,150,505	15,558	3,375,781
Total	5,790	341,228	56,740	20,090,287	62,530	20,431,400

Note Discrepancies between totals and actual sums are due to rounding errors.

Table J

Projected Numbers of Large Schools (Enrollments of 100 or More in Grades 1-6) and Their Grade 1-6 Enrollments, by Geographic Region and Poverty

Region	0-25%		26-35%		36-50%		51-75%		76-100%		Total	
	Schools	Enrollment	Schools	Enrollment	Schools	Enrollment	Schools	Enrollment	Schools	Enrollment	Schools	Enrollment
New England	2 619	768 485	361	105 001	344	106 653	248	64,131	160	52 199	3 731	1 096 468
Metropolitan Northeast	3 604	1 499 064	399	154 357	403	191 876	411	215 935	352	191 813	5 170	2 253 044
Mid Atlantic	3 614	1 276 353	942	324 650	764	264 558	665	261 415	252	107 511	6 237	2 234 487
Southeast	2 348	1 105 443	1 644	653 919	1 909	702 564	1 754	652 420	1 137	392 908	8 793	3 507 254
North Midwest	10 041	3 204 122	1 190	356 742	885	273 248	429	164 922	462	252 779	13 007	4 251 812
South Central	2 358	890 219	901	302 009	1 213	395 359	986	361 626	673	257 573	6 131	2 206 786
Central Midwest	2 819	799 187	405	88 808	262	58 046	142	42,235	171	29 573	3 738	1 017 846
North Central	1 681	507 168	208	55 416	71	17 640	56	12 961	58	13 043	2 075	606 229
Pacific Southwest	3 631	1 410 601	630	243 197	502	201 680	632	222 524	331	160 [illegible]82	5 727	2 238 181
Pacific Northwest	1 729	540 682	212	74 029	125	39 602	10	5 981	59	18 597	2 135	678 889
Total	34 445	1[illegible]001 356	6 893	2 358 121	6 478	2 251 214	5 332	2 004 137	3 595	1 476 179	56 743	20 091 006

Note: The discrepancies between totals and the actual sums are due to rounding errors

Table K

Projected Numbers of Large Schools (Enrollments of 100 or More in Grades 1-6) and Their Grade 1-6 Enrollments, by Urbanism and Poverty

Urbanism	Poverty (Free Lunch) Concentration											
	0-25%		26-35%		36-50%		51-75%		76-100%		Total	
	Schools	Enrollment	Schools	Enrollment	Schools	Enrollment	Schools	Enrollment	Schools	Enrollment	Schools	Enrollment
City of over 500 000	1 302	601 921	441	216 735	589	310 744	827	438 556	1 352	741 075	4 511	2 309 030
City of over 200 000-500 000	1 104	482 452	277	114 923	234	100 285	465	203 095	287	103 088	2 368	953,843
City of 50 000-200 000	3 659	1 302 978	829	301 631	935	348 261	750	285 759	383	139 144	6 555	2 377 770
Suburb of City	7 904	3 211 761	558	210 423	305	137 456	285	102 299	53	19 328	9 105	3 681 271
City under 50 000	10 605	3 540 337	2 290	794 173	2 092	699 281	1 283	478 014	563	180 782	16 834	5,692,588
Rural area near city	4 218	1 406 505	687	204 071	584	184 464	358	106 424	81	24 216	5 928	1 925,681
Rural area not near city	5 652	1 505 235	1 810	516 157	1 739	470 733	1 365	390 001	877	268 542	11 442	3 150,666
Total	34 444	12 001 294	6 893	2 358 105	6 478	2 251 210	5 332	2 004 146	3 595	1 476 167	56 742	20,090,925

Note The discrepancies between totals and actual sums are due to rounding errors

CHAPTER 5.
THE EDUCATIONAL PROGRAMS OF CE AND NON-CE STUDENTS

Ming-mei Wang
Ralph Hoepfner
Moraye B. Bear
Gerrie Smith

Title I students are offered the same number of days of instruction, but their absenteeism is slightly higher than that for regular students, resulting in the fact that they receive about four fewer days of instruction per year. While Title I students are receiving their compensatory educational services, non-participating students receive other, but very similar services. When CE services are compared on the basis of the number of years children participate, Title I is found to have the greatest continuity. This may be because Title I is the most permanent of the CE programs, or because the most needy children are selected for it and they are the most likely to continue needing it.

The framers of the Title I laws and regulations have been very deliberate over the years not to specify what services would be delivered in the name of Title I. General requirements exist to focus the services on academic areas and to ensure that they are supplemental, but Congress has been loath to infringe much more on local decision-making. While refraining from interference in the content or process of CE, Congress did want information on some of the more mechanical aspects of the services that could be useful in framing regulations on comparability. Specifically, information was requested on length of school year, attendance rates, non-CE activities while CE participants receive their CE services, and continuity of CE services.

Length of School Year. One way that schools and districts can conform to the 'supplement not supplant' requirement of Title I would be to extend the school year for Title I participants (other ways might involve extending the length of the school day or furnishing summer-school instruction or improving the 'quality' of instruction). Average numbers of days in the school year are tabulated for the six CE participation groups in Table 5-1. The means are based on students in the first-year representative sample, and are based on students rather than schools in order to be consistent with attendance data to be discussed shortly. As can be seen, the means all hover close to 177 days, so it is clear that Title I students do not receive more days of schooling.

Attendance. Although the number of days that instruction is offered can be administratively controlled to some degree, the number of days each student receives instruction is much less controllable. Table 5-2 provides average attendance rates that would be helpful in determining whether Title I students receive the supplemental services they are offered. We can see that CE students, Title I students in particular, have slightly lower average attendance rates. A difference of about 2.5 percentage points is equal to about 4 days of schooling. We can also see that attendance is better in the higher grades.

Non-CE Activities. The 'supplement not supplant' requirement for Title I presents schools with a dilemma if they attempt to adhere to it too strictly. The problem, of course, is what to do with the non-Title I students while the Title I students are receiving their supplemental Title I services. Because schools have not yet adopted cryogenic techniques for controlling student activities, the non-Title I students invariably are doing something, and that something, if the school is being responsible to its charge, should be educative—but not something that the Title I services would be supplanting. Information on what non-Title I students do not only gives us data on how the 'supplement' is being implemented, it can also serve to help us better to understand the results of evaluation studies.

In Table 5-3 we can see that the most common alternative activities for non-participants are in the same subject area, giving a strong impression that supplanting is actually taking place. Less related activities, such as study, self-selected activities, library work, gym, and field trips are also common when considered as a class of activities not likely to give rise to accusations that Title I services are supplanting other services.

Continuity of CE Participation. Based on loud and public complaints by several large school districts that strict obedience to Title I rules for the selection of participants resulted in a revolving-door program for many marginal achievers, Congress also wanted information on the continuity of Title I for individual students. This information would be useful in considering changes to regulations that would encourage continuity instead of rigidly enforcing 'now-you-qualify, now-you-don't' rules.

Table 5-1

Mean Number of Days in the School Year, by CE Participation Status and Grade

CE Participation		Gr. 1	Gr. 2	Gr. 3	Gr. 4	Gr. 5	Gr. 6
Title I student in	Reading	176.7	175.8	176.8	177.8	178.2	177.5
Title I school	Math	177.0	174.6	176.6	177.7	177.4	175.9
Other CE student in	Reading	176.8	178.0	178.3	177.9	178.0	177.1
Title I school	Math	176.4	177.1	177.2	177.1	177.1	177.4
Other CE student in	Reading	178.6	177.7	178.0	176.8	177.4	177.0
Other CE school	Math	179.0	178.3	178.5	177.2	177.2	177.2
Non CE student in	Reading	176.2	177.0	177.5	177.4	177.8	176.9
Title I school	Math	176.3	177.0	177.6	177.5	178.1	177.1
Non CE student in	Reading	178.1	178.3	178.4	177.6	177.9	177.6
Other CE school	Math	178.1	178.2	178.3	177.5	177.9	177.5
Non CE student in	Reading	177.0	177.3	177.2	175.4	175.6	175.8
Non CE school	Math	177.0	177.3	177.2	175.4	175.6	175.8

Note: Standard deviations, sample sizes, totals, and means calculated to two decimal places can be found in Table A of the Appendix.

Table 5-2

Mean Attendance Rates, by CE Participation Status and Grade

CE Participation		Gr. 1	Gr. 2	Gr. 3	Gr. 4	Gr. 5	Gr. 6
Title I student in	Reading	93.0	94.6	94 1	94.6	94.4	94.9
Title I school	Math	91.6	93.5	93.8	94.3	94.2	94.1
Other-CE student in	Reading	93 8	94.6	94.7	95.2	94.1	95.0
Title I school	Math	94 1	95.0	94.3	95 1	94.2	95.5
Other-CE student in	Reading	94.6	95.8	95 9	95.4	95.6	94.6
Other-CE school	Math	94.[illegible]	96.1	95.8	95.5	95.2	94.1
Non-CE student in	Reading	94 7	95.7	95 8	95.9	95.9	96 0
Title I school	Math	94 7	95.6	95.6	95.8	95.7	96.0
Non-CE student in	Reading	96 1	96.7	96.8	96.4	96.6	96.0
Other-CE school	Math	96.0	96.6	96.7	96.3	96.6	96.0
Non-CE student in	Reading	95 7	96.4	97.0	96.6	96.5	96.1
Non-CE school	Math	95.8	96.4	97 0	96.6	96.5	96.1

Note Standard deviations, sample sizes, totals, and means calculated to two decimal places can be found in Table B of the Appendix

Table 5-3

Percentage of Teachers in Title I Schools Who Reported That Non-Participants Are Involved in Alternative School Activities While Their CE Students Participate in CE Activities

Type of Activities in Which Non-Participants Are Involved	Reading Teachers for Reading Activities	Math Teachers for Math Activities
Reading or language arts activities	89 2	34.5
Math activities	29 8	81.9
Other subject matter activities	42.8	35.1
Study time	31.2	30.2
Student selected activities	27.4	26.8
Visits to the school library	16.3	10.2
Physical education activities	4 6	4.3
Field trips	1.8	0.9

Note Numbers supporting the percentages, separately by grade, and other relevant statistics can be found in Tables C and D of the Appendix.

We had data on CE participation for two years at the time Congress requested the information (see Kenoyer, Cooper, Saxton, and Hoepfner, 1980 for a comprehensive discussion of the data) For the first year (1975-76) we merely had retrospective reports of whether the student participated in any CE reading or math programs. In the 1976-77 school year our information included whether the CE was Title I or some other kind. We also collected the participation information from two sources. the primary source being a school coordinator who was to consult school or district records, and the secondary source being the classroom teachers who simply reported their knowledge (perceptions) about each student's CE participation (see Hemenway, Wang, Kenoyer, Hoepfner, Bear, and Smith, 1978, for a discussion of the problems in obtaining this seemingly simple information).

Tables 5-4 and 5-5 present information on CE continuation in two different ways. In Table 5-4 we categorized students by their current CE statuses, and then tabulated how many participated in the previous year. Because reading CE is more common than math, the percentages are always higher for reading. Among the three groups of CE participation, Title I students have the highest rate of previous participation. (This finding could be the result of numerous causes—the transitory nature of many other-CE programs, the lack of success of Title I in effecting achievement growth that results in a student's disqualification, the fact that Title I participants are the lowest of the low achievers, etc.) The converse is presented in Table 5-5, where all the CE participants in the first year are tabulated by their current-year status The percentages are highest for Title I, probably because it is the largest CE program.

We were also interested in learning the extent of agreement between our two sources of information on CE participation, and so calculated the percentage of teachers' reports (secondary sources) that agreed with our primary sources. The agreement rates, presented

Table 5-4

Percentage of Students Who Received CE in the Previous Year, by Current CE Status and Grade

CE Participation		Gr. 1	Gr. 2	Gr. 3	Gr. 4	Gr. 5	Gr. 6
Title I participants in 1976-77 who received CE in 1975-76	Reading	23.3	50 2	62 3	64.1	67.8	67.2
	Math	26.6	38.6	47 4	47 0	56.2	58 3
Other-CE participants in 1976-77 who received CE in 1975-76	Reading	27 6	35.8	42 1	51.0	46 6	47 0
	Math	29.2	33.6	23 7	34 7	31.5	30.9
Non-CE students in 1976-77 who received CE in 1975-76	Reading	3 3	6 4	9.4	8.9	8.5	6.5
	Math	2.2	2 7	3 8	4.5	4.9	4.4

Note Sample sizes on which the percentages are based can be found in Table E in the Appendix.

Table 5-5

Percentage of Students Who Received CE in the Current Year, by CE Status in the Previous Year and Grade

CE Participation		Gr. 1	Gr. 2	Gr. 3	Gr. 4	Gr. 5	Gr. 6
CE participants in 1975-76, who received Title I in 1976-77	Reading	33.4	43 7	43 9	40 0	40.6	35.3
	Math	27.2	35 1	41 7	33.4	36.2	29.8
CE participants in 1975-76 who received Other CE in 1976-77	Reading	33.7	25.7	21 4	25.0	23.6	27.6
	Math	39.2	32 7	21 0	28.1	22.6	25.0
CE participants in 1975-76 who no longer received CE in 1976-77	Reading	32 9	30.7	34 6	35.0	35.8	37.1
	Math	33 5	32.1	37.3	38.5	41 3	45.2

Note Discrepancies are due to rounding errors Sample sizes on which the percentages are based can be found in Table F of the Appendix

in Table 5-6, are surprisingly low, especially for other CE. We expected that other CE would have the lowest agreement, simply because the programs are generally not as highly regulated and controlled as Title I. These findings are cause for serious concern: how can teachers effectively deliver supplementary CE services to students if they don't know which students are participating?

Table 5-6

Percentage of Teachers' Reports of CE Participation That Agree With SES CE Participation Status

CE Participation Status	Gr. 1	Gr. 2	Gr. 3	Gr. 4	Gr. 5	Gr. 6
Reading						
Title I	62.1	78 3	77 7	79.8	80.5	79.4
Other-CE	41.0	47.1	43 0	47.7	49.3	47.3
No CE	95 0	94.1	94.2	94.3	94 8	94.6
Math						
Title I	58 5	69.8	70.0	71 1	77 3	74.1
Other-CE	36 8	34 1	29.3	35.2	34.9	32.2
No CE	96 7	96 6	96.9	96 4	96.2	97.0

Note Numbers on which the percentages are based and explanations of the data sources can be found in Table G of the Appendix

APPENDIX

SUPPLEMENTAL TABLES FOR CHAPTER 5

Table A

Length of School Year, by Reading and Math CE Participation and Grade

Grade	In Title I Schools						In Schools with Only Other CE									In Schools with No CE		
	Title I Students			Other-CE Students			Non-CE Students			Other-CE Students			Non-CE Students			All Students		
	Mean	S.D.	N	Mean	S.D.	N	Mean	S.D.	N	Mean	S.D.	N	Mean	S.D.	N	Mean	S.D.	N
Reading																		
1	176 71	4 80	1,602	176 77	4 73	815	176 21	6 53	6,613	178 58	3 50	533	178 11	5 00	2,786	177 02	5 33	1,268
2	175 77	6 48	1,762	178 02	2 98	826	176 96	5 62	5 462	177 70	3 92	588	178 34	4 85	2,441	177 26	5 10	1,111
3	176 84	5 47	1 880	178 33	2 91	706	177 48	5 03	5,278	177 97	4 16	558	178 38	4 87	2,489	177 15	4 96	1,160
4	177 78	5 46	1,540	177 94	3 06	682	177 37	4 93	5,434	176 83	3 43	527	177 64	4 14	2,487	175 36	6 88	1,184
5	178 24	5 41	1 423	177 98	3 35	640	177 83	5 14	5,622	177 41	4 10	542	177 87	3 96	2,640	175 59	6 73	1,250
6	177 51	5 52	1 163	177 08	3 58	724	176 94	5 18	5,962	176 96	3 60	541	177 57	3 54	4,334	175 84	6 03	1,662
Total	177 07	5 62	9,370	177 66	3 56	4,393	177 10	5 50	34,371	177 58	3 85	3,289	177 94	4 36	17,177	176 33	5 96	7,635
Math																		
1	177 04	4 45	850	176 38	4 45	731	176.26	6 41	7 411	179 00	3 74	361	178 09	4 90	2,957	177 02	5 34	1,264
2	174 60	6 90	858	177 06	3 09	626	177 05	5 59	6 545	178 34	3 10	306	178 20	4 84	2,716	177 30	5 07	1,104
3	176 63	5 15	1 006	177 16	3 23	532	177 55	5 09	6,322	178 46	3.56	389	178 28	4 90	2,657	177 15	4 97	1,158
4	177 72	5 19	895	177 14	3 54	626	177 49	4 99	6,124	177 21	2 63	362	177 53	4 20	2,646	175 35	6 88	1,183
5	177 38	6 23	819	177 11	3 77	539	178 05	4 99	6,326	177,21	3 74	404	177 88	4 02	2,773	175 59	6 74	1,247
6	175 89	5 80	657	177 35	3 28	634	177 11	5 16	6,546	177 17	2 82	424	177 54	3 61	4,440	175 84	6 03	1,659
Total	176 57	5 74	5,085	177 01	3 63	3,688	177 22	5 46	39,274	177 86	3 38	2,246	177 89	4 38	18,189	176 33	5 96	7,615

Table B

Rates of Attendance, by Reading and Math CE Participation and Grade

Grade	In Title I Schools						In Schools with Only Other CE									In Schools with No CE		
	Title I Students			Other-CE Students			Non-CE Students			Other-CE Students			Non-CE Students			All Students		
	Mean	S.D.	N	Mean	S.D.	N	Mean	S.D.	N	Mean	S.D.	N	Mean	S.D.	N	Mean	S.D.	N
									Reading									
1	92 95	7 98	1,572	93 78	7 38	796	94,72	6.81	6,252	94 59	6 44	521	96 08	4.95	2,685	95 74	6.15	1,201
2	94 55	6 65	1,741	94 63	6 32	810	95 66	5 77	5,166	95 82	5 63	575	96,69	4 26	2,341	96 36	5 02	1,056
3	94 11	7 14	1,847	94 71	6 66	689	95 80	5 95	4,997	95 90	5 11	530	96 76	4 57	2,446	96 96	4.64	1,106
4	94 56	7 28	1,524	95 23	6 37	675	95 91	6 42	5,172	95 35	5 82	520	96 41	5 07	2,414	96.62	4 68	1,150
5	94 41	7 91	1,403	94 14	7 83	623	95 86	6 55	5,351	95 56	5 79	540	96 63	4 85	2,562	96 53	5 05	1,224
6	94 89	7 22	1,146	95 02	7 03	715	96 04	6 09	5,769	94 61	6 64	539	96 02	5 74	4,259	96 05	5 78	1,632
Total	94 21	7 38	9,233	94 57	6 95	4,308	95 64	6 31	32,707	95 31	5 94	3,225	96 38	5 03	16,707	96 35	5 31	7,369
									Math									
1	91 57	8 91	829	94 10	6 99	710	94 70	6 75	7,061	94.93	6 30	349	95 96	5 08	2,857	95 78	6 08	1,198
2	93 52	7 04	842	95 02	5 77	611	95 59	5 84	6,254	96 13	5 59	294	96 56	4 45	2,619	96 40	4 99	1,051
3	93 77	6 99	978	94 30	6 50	523	95 62	6 18	6,031	95 80	5 27	362	96 72	4 58	2,614	96 95	4.66	1,106
4	94 26	7 35	887	95 09	6 28	620	95 81	6 54	5,856	95 47	5 34	356	96 33	5 20	2,572	96 62	4 68	1,150
5	94 16	8 78	816	94 15	8 00	521	95 74	6 48	6,038	95 24	6 09	403	96 62	4 84	2,697	96 54	5 05	1,222
6	94 10	7 75	651	95 55	6 81	625	96 01	6 03	6,347	94 08	7 63	422	96 04	5 61	4,371	96 06	5 77	1,631
Total	93 56	7 86	5,003	94 71	6 76	3,610	95 56	6 33	37,587	95 22	6 18	2,186	96 33	5 05	17,730	96 36	5 29	7,358

Table C

Percentage of Reading Teachers in Title I Schools Who Reported That Non-Participants Are Involved in Alternative School Activities While Their CE Students Participate in Compensatory Reading Activities

Type of Activities in Which Non-Participants are Involved**	Grade 1 (N = 272)		Grade 2 (N = 274)		Grade 3 (N = 268)		Grade 4 (N = 230)		Grade 5 (N = 201)		Grade 6 (N = 170)		Total (N = 1,415)	
	%	Count*	%	Count	%	Count	%	Count	%	Count	%	Count	%	Count
Other reading or language arts activities	96.0	261	89.8	246	89.6	240	80.9	186	84.1	169	77.7	132	89.2	1,234
Other reading or language arts activities only	36.8	100	30.7	84	29.9	80	24.8	57	35.3	71	29.4	50	31.2	442
Math activities	34.9	95	32.5	89	28.7	77	30.9	71	23.4	47	24.7	42	29.8	421
Activities related to other subject matter areas	29.8	81	40.5	111	43.3	116	49.6	114	48.3	97	51.2	87	42.8	606
Study time	22.1	60	29.6	81	34.3	92	36.5	84	31.3	63	35.9	61	31.2	441
Student selected activities	34.9	95	30.7	84	29.9	80	21.7	50	21.9	44	20.6	35	27.4	388
Visits to the school library	14.0	38	13.5	37	14.9	40	18.7	43	20.4	41	18.2	31	16.3	230
Physical education activities	2.6	7	5.5	15	5.2	14	6.1	14	3.5	7	4.7	8	4.6	65
Field trips	1.1	3	3.3	9	1.1	3	0.4	1	1.5	3	3.5	6	1.8	25

Table C (Continued)

	Response Pattern													
Reading teachers who have CE students	100.0	291	100 0	289	100.0	282	100.0	245	100.0	209	100.0	182	100.0	1,498
Who marked one or more of the above activities	93.5	272	94.8	274	95.0	268	93.9	230	96.2	201	93.4	170	94.5	1,415
Who marked none of the above activities	6.5	19	5.2	15	5.0	14	6.1	15	3.8	8	6.6	12	5.5	83

*Based on reading teachers in Title I schools who have CE students and who have reported at least one alternative activity for non-participants.

**Multiple responses were allowed.

Table D

Percentage of Math Teachers in Title I Schools Who Reported That Non-Participants Are Involved in Alternative School Activities While Their CE Students Participate in Compensatory Reading Activities

Type of Activities in Which Non-Participants are Involved**	Grade 1 (N = 137)		Grade 2 (N = 121)		Grade 3 (N = 136)		Grade 4 (N = 126)		Grade 5 (N = 118)		Grade 6 (N = 106)		Grade 7 (N = 744)	
	%	Count*	%	Count	%	Count	%	Count	%	Count	%	Count	%	Count
Other math activities	81.0	111	77.7	94	83.1	113	80.2	101	83.1	98	86.8	92	81.9	609
Other math activities only	36.5	50	31.4	38	27.2	37	36.5	46	34.7	41	47.2	50	35.2	262
Reading or other language arts activities	36.5	50	45.5	55	39.0	53	29.4	37	28.0	33	27.4	29	34.5	257

Table D (Continued)

Activities related to other subject matter areas	26.3	36	31.4	38	36.8	50	38.1	48	40.7	48	38.7	41	35.1	261
Study time	22.6	31	24.0	29	34.6	47	34.1	43	34.8	41	32.1	34	30.2	225
Student selected activities	29.9	41	27.3	33	34.6	47	27.8	35	21.2	25	17.0	18	26.8	199
Visits to the school library	4.4	6	7.4	9	11.8	16	12.7	16	14.4	17	11.3	12	10.2	76
Physical education activities	1.5	2	4.1	5	7.4	10	6.4	8	4.2	5	1.9	2	4.3	32
Field Trips	0.7	1	2.5	3	0.7	1	0.8	1	0.9	1	0.0	0	0.9	7
Response Pattern														
Math teachers who have CE students	100.0	176	100.0	156	100.0	170	100.0	167	100.0	141	100.0	125	100.0	935
Who marked one or more of the above activities	77.8	137	77.6	121	80.0	136	75.4	126	83.7	118	84.8	106	79.6	744
Who marked none of the above activities	22.2	39	22.4	35	20.0	34	24.6	41	16.3	23	15.2	19	20.4	191

*Based on math teachers who have CE students in Title I schools and who have reported at least one alternative activity for non-participants

**Multiple responses were allowed.

Table E

Students' Receipt of CE in 1975-76, by Their CE Participation Status in 1976-77*

CE Participation Status in 1976-77	Receipt of CE in 1975-76	Grade 1 Number	Grade 1 Percent	Grade 2 Number	Grade 2 Percent	Grade 3 Number	Grade 3 Percent	Grade 4 Number	Grade 4 Percent	Grade 5 Number	Grade 5 Percent	Grade 6 Number	Grade 6 Percent
Reading													
Title I students	Received	350	23 3	788	50 2	1,045	62 3	896	64 1	888	67 8	726	67 2
	Not Received	1 152	76 7	783	49 8	632	37 7	502	35 9	422	32 2	355	32 8
	Total	1,502	100.0	1 571	100 0	1,677	100 0	1 398	100 0	1 310	100 0	1,081	100 0
Other-CE students	Received	353	27 6	463	35 8	510	42 1	560	51 0	5[illegible]	46 6	566	47 0
	Not Received	926	72 4	830	64 2	702	57 9	538	49 0	591	53 4	639	53 0
	Total	1,279	100 0	1 293	100 0	1,212	100,0	1,098	100 0	1,106	100 0	1,205	100 0
Non-CE students	Received	345	3 3	553	6 4	823	9 4	783	8 9	783	8 5	762	6 5
	Not Received	10 163	96 7	8 147	93 6	7,934	90 6	8,039	91 1	8,427	91 5	10,944	93 5
	Total	10,508	100 0	8,700	100 0	8 757	100 0	8 822	100 0	9,210	100 0	11,706	100 0
Total Number of Students**		13,289		11 564		11 646		11,318		11,626		13,992	
Math													
Title I students	Received	209	26 6	295	38 6	410	47 4	367	47 0	422	56 2	357	58 3
	Not Received	577	73 4	469	61 4	455	52 6	414	53 0	329	43 8	255	41 7
	Total	786	100 0	764	100 0	865	100 0	781	100 0	751	100 0	612	100 0
Other-CE students	Received	301	29 2	275	33 6	206	23 7	309	34 7	263	31 5	299	30 9
	Not Received	729	70 8	544	66 4	663	76 3	581	65 3	571	68 5	669	69 1
	Total	1,030	100 0	819	100 0	869	100 0	890	100 0	834	100 0	968	100 0
Non-CE students	Received	257	2 2	270	2 7	367	3 8	424	4 5	481	4 9	542	4 4
	Not Received	11 207	97 8	9,657	97 3	9,411	96 2	9,067	95 5	9,406	95 1	11,760	95 6
	Total	11 464	100 0	9,927	100 0	9 778	100 0	9 491	100 0	9,887	100 0	12,302	100 0
Total Number of Students**		13 280		11,510		11,512		11,162		[illegible],472		13,882	

*Receipt of CE in 1975-76 was based on teacher's report in SBC CE participation status in 1976-77 was based on CER

**Students whose teachers indicated no records available for them in 1975-76 were excluded Please note that for students in Grade 1, CE receipt in 1975-76 referred to their receipt in Kindergarten

Table F

CE Participation Status for Students Who Received CE in 1975-76*

Grade		Title I Students	Other-CE Students	Non-CE Students	Total
		Reading			
1	Number	350	353	345	1,048
	Percent	33.4	33.7	32.9	100.0
2	Number	788	463	553	1,804
	Percent	43.7	25.7	30.7	100.1**
3	Number	1,045	510	823	2,378
	Percent	43.9	21.4	34.6	99.9**
4	Number	896	560	783	2,239
	Percent	40.0	25.0	35.0	100.0
5	Number	888	515	783	2,186
	Percent	40.6	23.6	35.8	100.0
6	Number	726	566	762	2,054
	Percent	35.3	27.6	37.1	100.0
		Math			
1	Number	209	301	257	767
	Percent	27.2	39.2	33.5	99.9**
2	Number	295	275	270	840
	Percent	35.1	32.7	32.1	99.9**
3	Number	410	206	367	983
	Percent	41.7	21.0	37.3	100.0
4	Number	367	309	424	1,100
	Percent	33.4	28.1	38.5	100.0
5	Number	422	263	481	1,166
	Percent	36.2	22.6	41.3	100.1**
6	Number	357	299	542	1,198
	Percent	29.8	25.0	45.2	100.0

*Receipt of CE in 1975-76 is based on teachers' report in SBC.
CE participation status in 1976-77 is based on CER.

**Discrepancies due to rounding errors.

104

Table G

Students' CE Participation Status in 1976-77 and Teachers' Reports of Their CE Receipt in 1976-77*

Grade		Title I Students		Other-CE Students		Non-CE Students		Total	
		Number	Percent	Number	Percent	Number	Percent	Number	Percent
Reading									
1	Receiving	1,045	7.5	570	4.1	540	3.9	2,155	15.5
	Not Receiving	638	4.6	819	5.9	10,335	74.1	11,792	84.6
	Total	1,683	12.1	1,389	10.0	10,875	78.0	13,947	100.1
2	Receiving	1,418	11.4	682	5.5	538	4.3	2,638	21.3
	Not Receiving	388	3.1	766	6.2	8,601	69.4	9,755	78.7
	Total	1,806	14.6	1,448	11.7	9,139	73.7	12,393	100.0
3	Receiving	1,480	12.0	557	4.5	533	4.3	2,570	20.8
	Not Receiving	425	3.4	737	6.0	8,625	69.8	9,787	79.2
	Total	1,905	15.4	1,294	10.5	9,158	74.1	12,357	100.0
4	Receiving	1,233	10.4	572	4.8	526	4.4	2,331	19.6
	Not Receiving	312	2.6	628	5.3	8,637	72.5	9,577	80.4
	Total	1,545	13.0	1,200	10.1	9,163	77.0	11,908	100.0
5	Receiving	1,177	9.6	588	4.8	493	4.0	2,258	18.5
	Not Receiving	286	2.3	605	5.0	9,079	74.3	9,970	81.5
	Total	1,463	12.0	1,193	9.8	9,572	78.3	12,228	100.0
6	Receiving	931	6.4	599	4.2	651	4.5	2,181	15.1
	Not Receiving	242	1.7	668	4.6	11,358	78.6	12,268	84.9
	Total	1,173	8.1	1,267	8.8	12,009	83.1	14,449	100.0
Math									
1	Receiving	515	3.7	409	2.9	391	2.8	1,315	9.5
	Not Receiving	366	2.6	703	5.1	11,508	82.8	12,577	90.5
	Total	881	6.3	1,112	8.0	11,899	85.7	13,892	100.0
2	Receiving	600	4.9	316	2.6	352	2.9	1,268	10.4
	Not Receiving	260	2.1	610	5.0	10,113	82.6	10,983	89.7
	Total	860	7.0	926	7.6	10,465	85.4	12,251	100.1
3	Receiving	697	5.7	274	2.2	322	2.6	1,293	10.6
	Not Receiving	300	2.5	661	5.4	9,554	81.5	10,915	89.4
	Total	997	8.2	935	7.7	10,276	84.2	12,208	100.0
4	Receiving	626	5.3	339	2.9	357	3.0	1,322	11.2
	Not Receiving	255	2.2	625	5.3	9,559	81.3	10,439	88.8
	Total	881	7.5	964	8.2	9,916	84.3	11,761	100.0
5	Receiving	651	5.4	328	2.7	393	3.3	1,374	11.3
	Not Receiving	191	1.6	611	5.0	9,945	82.1	10,747	88.7
	Total	842	7.0	939	7.8	10,340	85.3	12,121	100.0
6	Receiving	488	3.4	334	2.3	377	2.6	1,199	8.4
	Not Receiving	171	1.2	703	4.9	12,229	85.5	13,103	91.6
	Total	659	4.6	1,037	7.3	12,606	88.1	14,302	100.0

*CE participation status was determined on the basis of CER filled by the school's coordinators in 1976-77. Teachers were also asked to indicate in the SBC whether their students were receiving CE in 1976-77.

CHAPTER 6. TITLE I STUDENTS AND SPECIAL SERVICES

Judith A. Hemenway and Ralph Hoepfner

In order to determine if the sample of Title I students in the Sustaining Effects Study (SES) contains a disproportional number of handicapped students, who might be expected to hold down achievement growth and thereby obscure measures of program effectiveness, tabulations of student characteristics were projected to the nation. The projections indicate that the Title I sample does not have a disproportional share of students with handicapping conditions.

Categorical government programs have usually been designed to serve specific target populations as focused responses to perceived needs. For Title I of ESEA, the target population is composed of low-achieving students in low-income neighborhoods. For P.L. 94-142 (Education for All Handicapped Children Act), the targets are students with handicaps that impede learning or participation in school activities. Although the words in the guidelines for each program are clear, the differences in the concepts defining the two target populations are not so clear. In other words, it is the cause of the low achievement that theoretically determines whether a low-achieving child is a Title I student (an educational or social cause) or a 94-142 student (medical or psychological cause). Because the cause is not always clear, especially when handicapping condition is debatably medical (e.g., learning disabilities, mild educable mental retardation, emotional disturbance), we can expect some students to be in the 'wrong' program, some in both programs, or some who, multiply eligible, are disqualified from each program on the basis of an assumed qualification for the other. Birman (1979) studied the incidences of these occurrences and concluded that duplication of services was not a major problem because teachers and administrators made sure that services to multiply-eligible students were coordinated. Birman was concerned with the duplication or denial of services, and concerned also to enlighten the processes of making law and regulations so that such problems would be resolved.

Complementary to Birman's concerns, interest in the issue of multiply-eligible students and their services in the SES was not directed at duplication of services, but at multiple participation. Receipt of Birman's report in Washington caused questions regarding just how prevalent it was that Title I participants were also handicapped. The concern for the SES was that if special-education children, who might have low probabilities of normal achievement growth, are counted as Title I students, they may depress the measures of growth for the Title I samples. If such students are represented more in the Title I sample than in any comparison sample, and if their growth rates in achievement cannot be expected to be the same as that for other children in the Title I sample, then comparisons between samples may unfairly reflect poorly on the Title I program.

INFORMATION ON STUDENTS WITH PROBLEMS

An instrument in the study requested information relative to the issue of the incidence of handicapped students among Title I participants. Specifically, one item of the Student

Background Checklist asks for teachers' observations and judgments whether the student has physical, psychological, or adjustment problems that interfere with academic performance." It should be noted that the item posed no requirement for formal diagnosis by a physician, psychologist, speech therapist, etc. The responses were teachers' observations and judgments, and can be expected to reflect their attitudes and rationalizations to some unknown extent. National projections of the responses to this item were reported by Hemenway, Wang, Kenoyer, Hoepfner, Bear, and Smith (1978) in Table 4-7. Nationally, 11 percent of the students were judged by their teachers as having such problems. The high value of this percentage makes us wary of assuming that all, or even most, of these students are special, handicapped, or would qualify for services under P.L. 94-142.

From responses to a second item, we learned that of the 11 percent, 4.8 percent received treatment provided by the school, 1.7 percent received some treatment provided by the school but still had at least one untreated problem, and 4.5 percent received no treatment from the school. It should also be noted that there is no specification that the treatment provided by the schools met any current therapeutic standards.

In Table 6-1, we present a nationally projected cross-tabulation of the responses to this item with the teachers' judgments of each student's need for compensatory education (CE). The

Table 6-1

Teacher Judgments of Student Need for Compensatory Education and of Handicapping Conditions
(Entries are percentages of students, nationwide.)

Grade	Teacher Judgment of Student Need for CE	Student Has No Problems	School Gives Treatment	Some Treatment, Problems Remain	No Treatment
1	Does not need CE	95	3	0	2
	Needs CE	72	12	4	12
2	Does not need CE	95	2	1	2
	Needs CE	76	10	4	10
3	Does not need CE	95	2	1	2
	Needs CE	77	11	4	8
4	Does not need CE	94	3	1	2
	Needs CE	75	11	4	10
5	Does not need CE	96	1	1	2
	Needs CE	77	9	6	8
6	Does not need CE	96	1	1	2
	Needs CE	79	10	4	7

data in Table 6-1 are useful primarily as a validity check on the teachers' responses insofar as students with problems need extra services. In summary, about 95 percent of the students who 'do not need CE' have no problems that interfere with academic performance (the remaining 5 percent presumably have problems that do not interfere so much as to make them need CE services). More critically, about 24 percent of the students who 'need CE' are judged as having problems that interfere with academic performances (the remaining 76 percent presumably need it for other reasons). We can see from Table 6-1 that problems are much more common among the CE-needy students.

TITLE I STUDENTS AND ADDITIONAL CE SERVICES

Because special services to children with handicapping problems are frequently considered as compensatory services when the services are part of some program other than Title I, we can, to some degree, gauge how large the problem of the contamination of the Title I sample with special students might be by determining how many Title I students also participate in other CE programs. Table 6-2 provides national percentages of students by grade and participation in CE programs.

From Table 6-2 we can see that the percentages of children receiving more than Title I services is very small. This finding is similar, in some respects, to that of Birman.

In the lower half of Table 6-2, the children receiving CE services in addition to Title I are tabulated as percentages of the Title I population. These figures indicate that a substantial proportion of Title I students (about 14 percent) also receive other CE services. We interpret

Table 6-2

Percentages of Students in Title I Solely and With Other CE Services

Students in Title I Schools With:	Gr. 1	Gr. 2	Gr. 3	Gr. 4	Gr. 5	Gr. 6
National Percentages						
Reading						
Title I and Other CE	2.4	2.7	2.7	1.6	1.4	0.8
Title I only	10.4	13.7	13.0	13.0	10.8	8.7
Other CE only	6.4	7.1	5.9	6.3	5.6	5.7
Math						
Title I and Other CE	0.9	0.8	1.7	1.3	1.2	0.8
Title I only	5.5	7.3	6.6	6.8	6.7	5.5
Other CE only	7.2	6.9	5.4	6.6	5.5	4.8
Percentages of All Title I Students						
Reading	18.8	16.5	17.2	11.0	11.5	8.4
Math	14.1	9.9	20.5	16.0	15.2	12.7

these figures as extreme overestimates of the possible incidence of special students in Title I programs, because not all the additional CE services are of a special education nature. If the percentages were calculated on the basis of only those Title I students who are enrolled in schools that also participate in some other CE, so that it would be possible for them to be double participants, the projected percentages could be expected to be somewhat larger.

INCIDENCE OF CHILDREN IN EACH CE CATEGORY JUDGED TO HAVE PROBLEMS

We can look more specifically at children with problems and their frequencies in different CE programs. A first look at whether the incidence of problem students in Title I programs is greater than, equal to, or less than the incidences in other-CE categories that might serve as comparison groups in the SES is provided in Table 6-3. The percentages of children judged by their teachers as having problems that interfere with academic performance are projected to the national population for each of the categories of CE participation.

From Table 6-3 we can see that the incidence of teachers' judgments of having problems is lowest for the Non-CE students, for both reading and for math. Considering reading CE participants only, the incidence of judged problems is highest for other-CE students in other-CE schools (a possible reflection of the special nature of those CE programs) and for the doubly served students in Title I.

Among the math students, the other-CE students in other-CE schools have the highest incidence of judged problems, except for Grade 5, where the doubly served Title I students have a high incidence. Overall, however, judgments of problems are not notably more common among the doubly-served Title I students. Such judgments are similar for almost all CE categories and grades. Therefore one can conclude that later comparisions of Title I students to other CE students will not be greatly influenced by differential incidences of problems.

INCIDENCE OF CHILDREN IN EACH CE CATEGORY WHO RECEIVE TREATMENT FOR THEIR PROBLEMS

Looking more closely at the children judged by their teachers to have problems that interfere with academic achievement, Table 6-4 presents percentages of those students who received treatment.

For grades 1 and 2, the doubly served Title I students have a higher incidence of treatment than any other category, but this is not the case in the higher grades, where other-CE students frequently have the highest incidence of treatment.

CONCLUSIONS

Teachers' judgments of their students' physical, psychological, or adjustment problems that interfere with academic performances are supported to some degree by their judgments of those students' CE needs. The number of Title I students who are also served by additional CE programs is about 14 percent. Those doubly served students frequently, but not always, have a higher incidence of judged problems, however, the incidence is not very dissimilar from that for the other categories of CE. Among the students with judged problems, the doubly served have a slightly higher incidence of receiving treatment from the school, but only at grades 1 and 2 The data presented do not argue, except very weakly, that the Title I program might suffer in comparison to other programs because of the inordinate incidence of handicapped students.

Table 6-3

Percentages of Students Judged To Have Problems That Interfere with Academic Achievement, Nationwide

CE Participation	Gr. 1	Gr. 2	Gr. 3	Gr. 4	Gr. 5	Gr. 6
Reading						
Students in Title I Schools						
Title I and other CE	18 2	18 2	16 2	31 5	35.0	29 7
Title I only	20 5	22 1	17 6	18 3	17 3	17 1
Other CE only	13 7	20 5	18 0	16 6	18 1	13 2
No CE	10 6	8 6	7 9	9.7	8 3	6 9
Students in Schools with Only Other CE						
Other CE	25 9	25 0	27 6	30.2	29 3	26 3
No CE	8 3	6.2	8 8	10.1	4 8	5 6
Students in Schools with No CE						
All students	11 4	8.8	10 0	9 5	11.3	8 1
Math						
Students in Title I Schools						
Title I and other CE	21 8	17 0	16 4	13 5	37 5	22 0
Title I only	18 7	19 1	19 4	18 0	16.0	12 9
Other CE only	12 3	15.5	14 1	17 9	18 7	13 3
No CE	11 8	11 5	9 6	11 1	9 2	8 3
Students in Schools with Only Other CE						
Other CE	24 8	22 6	23 6	30 0	27 7	25 3
No CE	9 8	8.8	10 4	10 7	6.0	6 2
Students in Schools with No CE						
All students	11 4	8.8	10 0	9 5	11 3	8 1

Table 6-4

Percentages of Students Who Received Treatment for Their Problems,* by Grade and by CE Category (Based on Nationally Projected Counts)

CE Participation	Gr. 1	Gr. 2	Gr. 3	Gr. 4	Gr. 5	Gr. 6
Reading						
Students in Title I Schools,						
Title I and other CE	78.6	70.3	69.1	58.7	73.1	65.3
Title I only	63.2	62.3	55.1	58.5	60.1	61.7
Other CE only	54.4	46.8	62.8	69.8	62.5	68.2
No CE	53.7	49.4	59.5	57.3	57.9	57.3
Students in Schools with Only Other CE						
Other CE	65.9	68.6	79.2	80.8	70.6	89.0
No CE	56.6	58.1	44.3	57.4	50.0	50.0
Students in Schools with No CE						
All students	58.3	44.3	56.0	50.0	58.0	52.4
Math						
Students in Title I Schools						
Title I and other CE	73.9	78.8	78.2	62.2	61.3	83.3
Title I only	67.9	54.9	45.4	72.6	46.9	89.2
Other CE only	63.4	51.6	62.4	45.8	74.3	69.2
No CE	54.6	54.7	61.1	59.4	59.8	53.0
Students in Schools with Only Other CE						
Other CE	62.9	59.5	76.2	87.3	76.5	89.8
No CE	59.1	64.1	52.9	57.0	50.8	54.8
Students in Schools with No CE						
All students	58.3	44.3	56.0	50.0	58.0	52.4

*Treatment was defined for this table as full or partial treatment provided by the school.

PART II. ACHIEVEMENT AND COMPENSATORY EDUCATION

Many inquiries during the 1978 reauthorization (and annually during appropriations hearings) concerned how and how well Title I services (and CE in general) are distributed at the student level. In the first chapter of this part, we develop several indexes of school-level 'targeting'—the extent to which low achievers are selected for CE participation. Some caution in the interpretation of the figures is warranted, however, since the application of any of our targeting indexes to all schools, without regard to the philosophies underlying their CE programs and the wide range of acceptable program structuring, results in some unfair evaluations of schools. The indexes are compared on the basis of their merits and shortcomings.

The second chapter addresses the use of teachers' judgments to select students for CE, a practice that is widespread but not well understood in terms of accuracy in comparison to scores on achievement tests. We wanted to demonstrate how much agreement there was, so that if agreement were high the less costly judgment could be encouraged as an acceptable method for selecting students. Judgments being what they are, we also studied what might influence them, so that, even if accurate, systematic biases or errors could be pointed out.

CHAPTER 7. TARGETING OF COMPENSATORY-EDUCATION SERVICES

Charles E. Kenoyer and Deborah M. Cooper

In this study we attempted to develop and test a method for calculating an index that would reflect how well compensatory education (CE) is being targeted to those students who need it. Several definitions of participation and of need for CE are studied, reflecting different notions of ideal allocations of services. Twenty-five indexes are discussed and evaluated; none can accommodate all the acceptable and reasonable variations in how CE services are allocated to students. Instead of recommending a universal index, then, we urge that when a targeting index is needed it be selected on an ad hoc basis to correspond with the rationale and implications of the program's laws or regulations, and that the user be fully aware of its limitations.

A study of the targeting of compensatory education was conducted as part of the Sustaining Effects Study (SES) in an attempt to determine the degree to which CE services are targeted to needy students. The purpose of the study was to develop an index to indicate matches between the need for and participation in CE programs, and to examine the merits of the index. The findings are intended to inform policy-makers and decision-makers regarding the usefulness of the derived targeting indexes, which might then be considered for use in studies of compliance.

This paper does not address CE targeting at high administrative levels, such as states and districts, but focuses on how well CE is targeted to children in schools that receive CE funds. The targeting indexes we report, therefore, assume all the prior allocations and are based on the within-school success in reaching students defined as needing extra services. A variety of indexes was developed, based on different concepts of what constitute acceptable or ideal allocations of CE services. We examined such a variety for two reasons. First, most CE programs (including Title I, which is the most carefully regulated) allow for some flexibility in how schools can select students to receive services. A range of selection procedures is allowed to reach the goal of serving the neediest students—narrowly constrained regulations will not generally work. Second, many schools offer more than one CE program, each allowing different student-selection procedures. We study the targeting of CE services while taking into consideration variant selection procedures. In this manner we consider various ways to define student needs (test scores, teachers' judgments), participation (yes/no, or intensity of service), and CE goals (all students should be needy; all non-CE students should not) and develop indexes based on them.

We have taken the position that no single index of targeting can serve for all situations in which targeting is to be assessed. What constitutes proper assignment of CE to students may be specified in various ways, depending on several considerations in addition to the regulations of the CE program. The selection of a targeting index for any given evaluation problem, then, requires that the evaluator exercise judgment. For this reason, with each index we present the rationale for its derivation and some empirical data on its distribution in the SES sample.

Due to the variety of situations, the evaluator must consider many things about the program before choosing a particular index to assess its targeting properties:

- How prescriptive are the regulations of the program in specifying which students are to participate? Can a school, under some circumstances, decide to select all students? Can it decide to limit the grades (and thereby the students) in which the program is to be implemented? Can it decide to continue services to students even though they are technically disqualified by other criteria?

- Should the index be sensitive to conditions in which funds are insufficient to meet school needs? For example, an index that considers only how many of the students are needy may be insensitive to funding limitations. But an index that considers how many of the needy students are to be selected for the program may be solely sensitive to the availability of funds.

- Are the allocation procedures based on national normative data?

- Should the targeting index consider only the selection of students, or should it consider the instructional services to be received by them? Selection is, in fact, little more than a designation that is part of the allocation procedures; it may not be a valid indicator of the receipt of any services likely to be called 'compensatory.'

- Should ease-of-computation of the index be considered. As will become apparent in this report, some indexes can be computed very simply, while others can be utilized only if one has access to a computer.

These indexes will be discussed in sequence, beginning with a relatively simple one and proceeding through progressively more complex and sophisticated ones.

Targeting of CE services can be traced, in theory, from the legislation that makes the funds available, through the relevant regulations on how the funds are to be allocated to states, districts, and then schools, up to the schools' allocations of services to students. (Other papers in this volume address the early steps in this process.) For the purposes of this study, only the last step in the allocation process, targeting within schools, will be considered. Because targeting at that level usually depends on academic needs, each targeting index will involve some measure of agreement between the selection of students for CE services (or their receipt of them) and a measure of the academic neediness of the students.

In the first section we consider different but related targeting indexes. These indexes represent the proportions of students selected for CE services who are in need of them. Need is defined with several cutoff points on the nationally normed score distributions of reading and math achievement tests. The proportions can range from 0.0 (no student selected for the CE services at or below the achievement-score cutoff) to 1.00 (all selected students at or below the cutoff). These indexes are based on national definitions of need (the cutoffs) and not on local school definitions. They do not consider the neediness of students *not selected* for CE.

Indexes based on correlations between measures of need for CE and measures of selection for CE (or receipt of instructional services) are investigated in the second section in an effort to consider the needs of both CE and non-CE students. Indexes for schools are always more restricted in range than those computed for grades, and indexes for reading are always higher than those for math. Although few of the values of the various indexes based on coefficients are negative (indicative of gross mis-targeting), few of them are positive and sufficiently large enough for statistical significance at the 0.5 level (indicative of non-chance targeting). When need for CE is defined by teachers' judgments instead of test scores, the values are much higher, but such indexes are confounded by a lack of independence between the judgment and the teacher's knowledge of each student's CE status. On logical grounds, adjusted correlation coefficients ought to be more valid as targeting indexes; when computed they also tend to be higher in value.

When measures of the amounts of instructional services received are used instead of a mere dichotomous indicator of CE participation, the values of the targeting indexes are much lower. This is caused by the consideration of all schools, even those without CE (likely to be minimal, especially for reading, because almost all schools had CE students for reading), and by the fact that amounts of services received by CE and non-CE students overlap a great deal.

An even greater refinement of targeting indexes may be found in the third section, where the dichotomous measures of student need for services (the use of cutoffs or teachers' judgments of need or no need) are replaced by a continous scale of percentile scores. The improvement in the size of indexes, however, is little more than one could expect merely statistically from coefficients based on a finer scale of measurement. The problem remains, however, that the use of a nationally-normed metric of need simply cannot accurately reflect the degree of targeting within a school, when each school must target its services on the basis of need relative to its student body.

To take this relative need into account, the fourth section develops some additional indexes based on the achievement ranks of students within schools. Several related indexes are investigated that yield different values. A correlational approach relating relative need ranking to CE participation appears to offer the most sensible and usable index of targeting. From these indexes we can conclude that about 75 percent of the schools target reading CE well beyond chance levels, and about 50 percent target math CE as adequately. When grades are considered, however, less than one-third of the first grades are well targeted for reading or math. At the higher grades, about 75 percent target reading CE well and about 60 percent target math well.

INDEXES BASED ON THE PROPORTION OF CE PARTICIPANTS WHO ARE 'NEEDY'

The first approach to a targeting index to be considered is based on the minimal, fundamental assumption of CE: namely, that all students being served are educationally in need of services. The most obvious index designed to reflect this is the proportion of CE students in a school who are educationally needy, where need is defined by scores on an achievement test. In cases where all participating students meet the criterion, the index value will be 1.0, when none of the students meet the criterion, the index value will be 0.0. The index ranges from zero to one, with intermediate values reflecting the degree to which targeting among the CE students is appropriate

The CTBS reading and math criterion scores were selected to reflect educational need in the calculation of these proportions. Three different cutoff scores were selected, and anlyses performed in which students were classified as needy or not needy on the basis of each.

Below-Median Scores as the Definition of Need. The first and highest of the cutoff points is the 50th percentile, the median. This score represents an extremely high cutoff for the definition of need. We would expect to find agreement that children scoring at or above a median based on national norms belong to the 'not needy' category, but it is not so obvious that the cutting score is low enough—it is not likely that we would find agreement that all students scoring below the median are 'needy.' (Moreover, it will not likely ever be economically feasible to provide costly CE programs to one half of the student population.) Nonetheless, we start with the median to provide a point of reference. CE students for reading and math were classified as needy if they scored below the national median, and as not needy if they scored at or above it. The number of CE students at each school scoring below the national median was then divided by the total number of CE students at that school, to generate the index for the school. The same procedure was repeated for each grade. Cumulative frequency distributions of the index for schools and for grades are presented in Table 7-1 for reading and for math. The indexes were based on data from each school in the SES first-year sample (for the school-level indexes) or each grade (for the grade-level indexes) in which there were CE students. Notice that the intervals of index values in Table 7-1 (and other tables) stop at 0.96. By subtracting the cumulative percentage at that value from 100.0, the reader can quickly determine the percentage of schools (or grades) that exhibit perfect or near perfect targeting.

For reading, fewer than 3 percent of the schools had index values (first column in Table 7-1) less than or equal to .50 (indicating that at least half of the CE students were not needy by our critieron), and almost 15 percent scored at .96 or more (indicating almost perfect targeting). For math, the corresponding cumulative percentages of schools were generally higher, which indicate that targeting for math CE is poorer (except that more schools seem to have near perfect targeting for math). For both reading and math, the first grade shows higher percentages of schools in the lower range, especially near zero, than any other grade, probably reflecting low test validity at that grade.

The 40th and 35th Percentiles as Definitions of Need. Additional cutoff points, lower than the median, were also examined with the hope that one of them might meet with wider agreement as a meaningful or useful definition of educational need. A guideline can be obtained from another report (Kenoyer, "Teacher Judgment of Need for Compensatory Education," this volume), however, in which we find that the cutting score that minimized the disagreement between teacher classification and test classification varied somewhat

Table 7-1

Cumulative Percentage Distributions by School and Grade: Proportions of Students in Reading CE and Math CE Who Score Below the 50th Percentile in Those Subjects

Index Value	School (N = 206)	Grade 1 (N = 149)	Grade 2 (N = 176)	Grade 3 (N = 180)	Grade 4 (N = 169)	Grade 5 (N = 160)	Grade 6 (N = 146)
Reading							
0.00-0.08	0.0	6.0	1.1	0.6	1.8	0.6	0.7
0.09-0.16	0.0	6.7	1.1	0.6	1.8	0.6	0.7
0.17-0.24	0.0	8.7	1.7	1.1	1.8	0.6	0.7
0.25-0.32	0.0	10.1	2.8	1.1	1.8	1.2	0.7
0.33-0.40	0.5	14.8	4.0	1.7	5.9	1.9	1.4
0.41-0.48	1.5	16.8	4.5	2.2	6.5	1.9	1.4
0.48-0.56	3.4	28.2	10.2	4.4	8.3	5.0	4.1
0.57-0.64	5.8	43.0	14.2	8.9	11.8	7.5	4.8
0.65-0.72	17.0	61.1	18.8	12.8	14.8	13.1	9.6
0.73-0.80	28.6	77.2	27.3	22.2	22.5	19.4	17.8
0.81-0.88	52.4	88.6	41.5	37.2	34.3	31.3	29.5
0.89-0.96	85.4	91.3	54.5	49.4	48.5	45.0	42.5
Math							
Index Value	**School (N = 161)**	**Grade 1 (N = 90)**	**Grade 2 (N = 100)**	**Grade 3 (N = 114)**	**Grade 4 (N = 113)**	**Grade 5 (N = 112)**	**Grade 6 (N = 107)**
0.00-0.08	0.0	4.4	3.0	2.6	0.0	0.0	1.9
0.09-0.16	0.0	4.4	3.0	4.4	0.9	0.0	0.0
0.17-0.24	0.0	6.7	4.0	4.4	1.8	0.0	0.0
0.25-0.32	1.2	8.9	5.0	5.3	5.3	0.0	0.0
0.33-0.40	3.1	12.2	8.0	7.0	7.1	0.0	2.8
0.41-0.48	5.6	14.4	9.0	7.0	7.1	2.7	4.7
0.49-0.56	9.9	30.0	18.0	15.8	13.3	5.4	5.6
0.57-0.64	14.9	41.1	23.0	16.7	21.2	10.7	7.5
0.65-0.72	25.5	53.3	30.0	25.4	28.3	16.1	19.6
0.73-0.80	38.5	61.1	47.0	30.7	38.9	30.4	30.8
0.81-0.88	62.1	70.0	57.0	43.0	53.1	42.9	42.1
0.89-0.96	80.7	80.0	61.0	58.8	63.7	52.7	55.1

between reading and math and over grades, but was bracketed fairly well by the 35th and 40th percentile. The 40th and 35th percentiles were therefore used as definitions of educational need in the calculation of targeting indexes. Tables 7-2 and 7-3 contain the distributions based on dichotomizations at the 40th percentile and the 35th percentile, respectively. Index values based on these cutoffs are, of course, lower than those based on the 50th percentile, because the criterion is more stringent (a greater proportion of CE students must come from the lower achievement levels). Using the 35th percentile, for example, only 1 percent of the schools exhibit perfect or near perfect (.97 or higher) targeting for reading, and about 9 percent do so for math (whereas for the median as cutoff,

Table 7-2

Cumulative Percentage Distributions by School and Grade: Proportions of Students in Reading CE and Math CE Who Score Below the 40th Percentile in Those Subjects

Reading							
Index Value	School (N = 206)	Grade 1 (N = 149)	Grade 2 (N = 176)	Grade 3 (N = 180)	Grade 4 (N = 169)	Grade 5 (N = 160)	Grade 6 (N = 146)
0.00-0.08	0.0	7.4	2 3	1.1	3.6	1.2	2.1
0.09-0.16	0.0	8.1	3.4	1.7	4.1	1.2	2.1
0.17-0.24	0 0	10.1	5.1	2.2	4.7	1.2	2.7
0.25-0 32	1 5	13.4	6 3	3.9	6.5	1.9	2.7
0 33-0 40	2 4	21 5	10 2	7.8	9 5	3.7	5.5
0.41-0 48	6 3	28 9	15.3	8.9	10.7	5 6	6.8
0 49-0 56	12 6	44 3	24 4	16 1	14.2	13.7	12.3
0 57-0 64	24 3	63.1	31 8	21 7	20.1	19.4	15.1
0.65-0 72	37 4	79.2	38.6	31 7	28.4	28.1	26.0
0 73-0.80	56 8	89 3	56 8	43.9	39.6	40.0	40.4
0 81-0 88	81 1	96 0	71 6	55.0	56 8	53.7	58.9
0 89-0.96	95.1	96.6	80 1	68 9	66.3	64.4	66.4

Math							
Index Value	School (N = 161)	Grade 1 (N = 90)	Grade 2 (N = 100)	Grade 3 (N = 114)	Grade 4 (N = 113)	Grade 5 (N = 112)	Grade 6 (N = 107)
0 00-0 08	0.0	6.7	5 0	5 3	0.9	2.7	2.8
0 09-0 16	0 0	7 8	6.0	7.0	1.8	3.6	2.8
0 17-0.24	0 6	11.1	8.0	7 9	4.4	4.5	2.8
0 25-0 32	5 0	15.6	12.0	10.5	6.2	4.5	4.7
0 33-0.40	7 5	20 0	15 0	13.2	11.5	5.4	9.3
0.41-0 48	13.0	23.3	18.0	16.7	12.4	7.1	11.2
0 49-0 56	18 0	43.3	31.0	18.4	25.7	17.9	18.7
0 57-0.64	26 1	54.4	39 0	27.2	33.6	22.3	27.1
0.65-0.72	49.7	66.7	50.0	37.7	49.6	35.7	38.8
0 73-0.80	68 3	74.4	66.0	52 6	58.4	50.0	53 3
0.81-0.88	82.0	78 9	75.0	62.3	74.3	65.2	64.5
0 89-0.96	90 1	84.4	76.0	70.2	72.0	67.9	69.2

the percentages of schools doing that well are 15 and 19). The achievement-test definition of educational need one adopts clearly leads to different conclusions regarding how well schools are targeting CE.

A serious shortcoming of the indexes in this section is that they depend only on the students who participate in CE, and therefore a school having only low-achieving students can't help but have perfect targeting (an index of 1.0). Therefore, the indexes in the next section were developed using both CE and non-CE students.

113

Table 7-3

Cumulative Percentage Distributions by School and Grade: Proportions of Students in Reading CE and Math CE Who Score Below the 35th Percentile in Those Subjects

			Reading				
Index Value	School (N = 206)	Grade 1 (N = 149)	Grade 2 (N = 176)	Grade 3 (N = 180)	Grade 4 (N = 169)	Grade 5 (N = 160)	Grade 6 (N = 146)
0.00-0.08	0.0	11.4	3.4	1.7	4.1	1.9	2.7
0.09-0.16	0.0	13.4	4.0	2.2	5.9	2.5	3.4
0.17-0.24	0.5	17.4	9.1	3.3	6.5	2.5	4.8
0.25-0.32	2.4	26.8	11.9	5.0	8.9	3.1	4.8
0.33-0.40	7.3	43.6	19.9	11.1	13.0	5.6	8.9
0.41-0.48	13.6	58.4	24.4	17.2	16.6	8.1	8.9
0.49-0.56	25.2	72.5	36.9	22.8	23.7	17.5	17.1
0.57-0.64	38.3	80.5	40.3	28.9	29.6	25.6	30.8
0.65-0.72	55.8	89.3	52.8	42.8	40.8	39.4	40.4
0.73-0.80	76.2	98.0	70.5	54.4	58.0	53.1	55.5
0.81-0.88	93.2	98.0	80.7	70.6	74.6	64.4	69.9
0.89-0.96	99.0	98.0	88.6	80.6	79.9	72.5	74.7
			Math				
Index Value	School (N = 161)	Grade 1 (N = 90)	Grade 2 (N = 100)	Grade 3 (N = 114)	Grade 4 (N = 113)	Grade 5 (N = 112)	Grade 6 (N = 107)
0.00-0.08	0.0	8.9	7.0	5.3	4.4	2.7	3.7
0.09-0.16	0.6	14.4	10.0	7.0	6.2	4.5	3.7
0.17-0.24	5.0	21.1	16.0	9.6	10.6	4.5	6.5
0.25-0.32	7.5	23.3	19.0	10.5	10.6	4.5	7.5
0.33-0.40	9.9	27.8	23.0	14.9	18.6	7.1	12.1
0.41-0.48	17.4	40.0	27.0	19.3	21.2	7.1	17.8
0.49-0.56	25.5	55.6	39.0	21.9	32.7	19.6	27.1
0.57-0.64	41.6	62.2	48.0	32.5	39.8	27.7	29.9
0.65-0.72	60.2	72.2	60.0	44.7	60.2	39.3	47.7
0.73-0.80	75.2	78.9	73.0	57.0	71.7	55.4	64.5
0.81-0.88	85.1	85.6	81.0	66.7	82.3	68.8	74.8
0.89-0.96	90.7	85.6	81.0	70.2	83.2	72.3	78.5

INDEXES BASED ON CORRELATIONS OF CE PARTICIPATION WITH NEED AND WITH SERVICES

One shortcoming of the indexes developed in the previous section is that they are unaffected by the incidence of educational need among the students not participating in CE Correlations that involve neediness and CE status, on the other hand, measure agreement between the two variables, with non-participants included as well as participants. In this section, dichotomous indicators of need are correlated with CE status first, then with measures of amount of instructional services received Where the dichotomous need indicators are correlated with CE status, itself a dichotomous variable, the resulting coefficient is a *phi* coefficient Where the need indicators are correlated with the quantitative variables indicating amount of service, there obtains a point-biserial coefficient.

Correlations with Indicators of Need

The refinement introduced with these additional indexes comes from consideration of students not selected for CE—students ignored in the calculation of the previous indexes. First, we consider a student's need for CE (or not) in terms of test scores and whether or not the student participates. Conceptualized in this way, the index will show a perfect value when all the CE students are needy (according to test score) and all non-CE students are not needy.

Unadjusted Phi Coefficient. The kind of targeting index to be considered first is the *phi* coefficient, the product-moment correlation between the dichotomous variable denoting each student's selection for reading CE (or math CE) and the dichotomous variable derived by cutting the student's CTBS reading (or math) score at one of the three levels used in the preceding section, i.e., the 50th, 40th, or 35th percentile. There is no *a priori* basis for knowing with certainty which of the three dichotomizations will yield the largest values for this index, since both CE and non-CE students are included in the calculation.

The theoretical range of the *phi* coefficient is from −1.0 to +1.0. Therefore, while the tables in this section are similar to earlier tables, the range of the index values is different. Table 7-4 contains the cumulative distribution of the index (*phi* coefficients) for schools and for each grade, for reading and math, where educational need is defined by dichotomizing the CTBS scale scores at the 50th percentile, based on national norms.

Looking first at the reading distributions, we find that nearly all values are positive, that only 1.5 percent of the indexes at the *school* level are negative, and that none are large. Within this range, the distribution is fairly symmetrical and bell-shaped, indicating a wide range of targeting accuracy. The distributions *by grade* have a wider range, with greater frequencies of extreme values in both directions. The difference in ranges is probably the result of the aggregation process that leads to greater reliability at the school level. Grade 1 has more low index values, with close to 20 percent of them less than zero. This is attributable, at least in part, to low test validity at this level, but is also probably influenced by lower accuracy of CE assignment from a relative unfamiliarity with student capabilities.

The distributions are a bit different for math, with the lower end of the range somewhat more negative at the school level. The same relative disadvantage in targeting at the first-grade level is apparent, and the index values by grade again tend to have greater scatter than the school-level values. This finding suggests that targeting is generally better at the school level because the student selection is less restricted. The school-level indexes, of course, are lowered when schools limit CE services to selected grades while there are many needy students in the excluded grades.

Table 7-4

Cumulative Percentage Distributions by School and Grade: *Phi* Coefficients Between CE Participation and Achievement Scores Dichotomized at the 50th Percentile

Index Value	School	Grade 1	Grade 2	Grade 3	Grade 4	Grade 5	Grade 6
Reading							
	(N = 205)	(N = 143)	(N = 168)	(N = 172)	(N = 143)	(N = 155)	(N = 144)
-.62 - -.53	0 0	0.0	0.0	0.0	0.0	0.0	0.7
-.52 - -.43	0 0	0.0	0.0	0.0	0.0	0.0	0.7
-.42 - -.33	0.0	1.4	0.0	0.0	0.0	0.0	0.7
-.32 - -.23	0 0	5.6	0.0	0.0	0.0	0.0	0.7
-.22 - -.13	0 0	8.4	0.6	1.2	1 2	0.0	1.4
-.12 - -.03	1 5	19.6	3.6	2.3	1 8	1.3	2.1
-.02-.02	2.0	25.2	4.2	2.9	3.1	2.6	2.8
.03-.12	7.3	51 7	11.9	5.2	9.2	7.1	4.9
.13-.22	26 8	72 0	22.6	19 2	22.1	18.1	23.6
.23-.32	50.2	88 8	33.9	33.7	37.4	33.5	50.7
.33-.42	78 5	94.4	48.8	47.7	57 7	54.8	66.7
.43-.52	95.1	97.9	66 1	68.6	75 5	74.2	81.9
.53-.62	99.0	98.6	77.4	87.2	92.6	87.7	94.4
.63-.72	99 5	98.6	89.9	92.4	98.1	95.5	97.2
.73-.82	100 0	99 3	95.8	97.7	99.4	98.1	97.9
.83-.92		99.3	98.8	98.2	99.4	98.7	98.6
Math							
	(N = 161)	(N = 86)	(N = 96)	(N = 109)	(N = 100)	(N = 107)	(N = 104)
-.62 - -.53	0.0	1.2	0 0	0.9	0.0	0.0	0.0
-.52 - -.43	0.0	1.2	0.0	0.9	0.0	0.0	0.0
-.42 - -.33	0 0	4.7	0.0	3.7	0.0	0 0	0.0
-.32 - -.23	0 0	4 7	1 0	4.6	0 9	0 0	0.0
-.22 - -.13	1.2	8.1	3.1	6.4	1.9	0.0	1.9
-.12 - -.03	6.8	19.8	9.4	9.2	5.6	0.0	3.8
-.02-.02	9.9	29.1	11.2	11 9	7.4	0.9	4.8
.03-.12	29.2	40.7	24.0	26 1	19.4	10.3	21.2
.13-.22	54.7	67.4	49.0	41.3	48.1	32.7	47.1
.23-.32	84 5	83.7	70.8	62.4	67.6	62.6	75.0
.33-.42	93.2	89.5	90.6	78.0	82.4	80.4	84.6
.43-.52	96.9	95.3	95.8	87.2	89.8	85.0	93.3
.52-.62	98.8	97.7	96.9	94.5	97.2	93.5	97.1
.63-.72	99.4	100.0	99.0	96.3	98.1	98.1	100.0
.73-.82	100.0		99.0	96.3	100.0	99.1	
.83-.92			100.0	97.2		99.1	

Note. The indexes in the middle set of rows of each subtable (-.22 to +.22) have values that are (approximately) not significantly different from zero—or chance targeting.

Table 7-5 contains the same distributional information, based on the 40th percentile dichotomization, and Table 7-6 is based on the 35th percentile cutoff. The three cutoff scores, when compared, do not lead to markedly different distributions. Summaries of the school-level distributions are shown in Table 7-7 for the purpose of comparison of the effects of the three dichotomizations. For both reading and math, the 35th and 40th percentile cutoffs yield higher index values than when the cut is at the median (50th percentile). Targeting of reading CE is clearly better than targeting of math CE, but there are many instances of ineffective or poor targeting in both cases.

Taking the median value for reading among the three cutoffs (i.e., the median for the *phi* based on the 40th percentile cut), we find that about 10 percent of the variance of the dichotomized achievement score and CE assignment is shared. The best math mean, based on the 35th percentile cut, indicates that about 4 percent of the variance is shared by the two variables. The implication is that this kind of index cannot account for the grade-by-grade and school-by-school variation in how students are selected for CE. Unless a strict and universal selection criterion were mandated, this kind of index is likely to indicate poor targeting.

Unadjusted Phi Coefficient with Teacher Judgment of Need for CE. In addition to the indexes above, all based on dichotomizations of the CTBS scores, a different index can be created by replacing the CTBS dichotomies with teacher judgment of each student's need for CE. A targeting index was derived for reading by correlating teacher judgment of need for reading CE (needs it or does not need it) with whether or not the student participates in reading CE, and similarly for math. In attempting to evaluate these indexes in relation to the CTBS indexes, it is important to be aware that in many cases the teachers probably knew the CE status of each student, thus contaminating their responses. Therefore, agreement between teacher judgment and CE status may be spuriously inflated, and these indexes must be interpreted carefully.

Table 7-8 contains the distributions of these indexes for reading and math for schools and for grades within schools. For both reading and math, the correlations tend to be more positive, for reading, there are more correlations greater than .80 than there are negative correlations in any grade. For math, the positive tendency is not as strong, but it too is clearly present. The teacher-judgment criterion of need for CE serves as a somewhat better correlate of CE status that the CTBS dichotomies, but it is impossible, given the circumstances of data collection, to rule out spurious inflation of the coefficients.

Adjusted Phi Coefficients. A problem with using the *phi* coefficient as a targeting index is that the theoretical range from −1.0 to +1.0 can be achieved only if the two dichotomous variables being correlated have equal marginal probabilities. For the present analyses, the departure from that condition can be expected to be severe, the number of students who can be selected for CE is influenced by a school's funding level, among other things, and the number of needy students is not under the control of the school (at least initially). Since our focus is upon the effectiveness of school targeting procedures, and because it seems appropriate to let the extraneous influences enter into the value of the targeting index, we should examine an adjusted coefficient.

Table 7-5

Cumulative Percentage Distributions by School and Grade: *Phi* Coefficients Between CE Participation and Achievement Scores Dichotomized at the 40th Percentile

Reading

Index Value	School (N = 205)	Grade 1 (N = 143)	Grade 2 (N = 167)	Grade 3 (N = 172)	Grade 4 (N = 163)	Grade 5 (N = 154)	Grade 6 (N = 144)
−.62 - −.53	0.0	0.7	0.0	0.0	0.0	0.0	0.0
−.52 - −.43	0.0	0.7	0.0	0.0	0.0	0.0	0.0
−.42 - −.33	0.0	2.1	0.0	0.0	0.0	0.0	0.0
−.32 - −.23	0.0	4.2	0.0	0.0	0.0	0.0	0.0
−.22 - −.13	0.0	9.1	0.6	0.6	0.6	1.3	0.7
−.12 - −.03	1.5	21.7	3.6	1.7	3.1	1.9	2.1
−.02-.02	1.5	31.5	5.4	2.9	4.9	3.9	4.2
.03-.12	6.3	53.8	14.4	6.4	8.0	5.8	7.6
.13-.22	20.5	72.7	20.4	14.5	19.6	14.3	24.3
.23-.32	46.8	88.1	29.9	25.0	34.4	28.6	41.7
.33-.42	76.6	93.7	48.5	45.3	51.5	43.5	61.1
.43-.52	94.6	97.9	67.1	63.4	72.4	67.5	79.9
.54-.62	99.5	97.9	82.0	83.7	85.3	83.1	90.3
.63-.72	100.0	99.3	92.8	91.3	95.1	92.9	97.2
.73-.82		99.3	98.8	96.5	98.1	96.8	100.0
.83-.92		99.3	100.0	98.8	98.8	98.1	

Math

Index Value	School (N = 161)	Grade 1 (N = 87)	Grade 2 (N = 98)	Grade 3 (N = 108)	Grade 4 (N = 108)	Grade 5 (N = 109)	Grade 6 (N = 104)
−.62 - −.53	0.0	0.0	0.0	0.0	0.0	0.0	0.0
−.52 - −.43	0.0	1.2	0.0	0.0	0.0	0.0	0.0
−.42 - −.33	0.0	3.5	0.0	0.9	0.0	0.0	0.0
−.32 - −.23	0.0	5.8	0.0	3.7	0.0	0.9	0.0
−.22 - −.13	0.6	8.1	2.1	3.7	0.9	0.9	1.9
−.12 - −.03	4.3	23.3	7.3	8.3	5.6	3.7	6.7
−.02-.02	7.5	30.2	15.6	12.0	8.3	4.7	10.5
.03-.12	27.3	41.9	29.2	18.5	23.1	13.1	23.8
.13-.22	55.3	67.4	47.9	36.1	47.2	29.9	44.8
.23-.32	82.6	83.7	65.6	56.5	67.6	53.3	71.4
.33-.42	93.8	93.0	80.2	77.8	81.5	75.7	82.9
.43-.52	98.1	95.3	93.8	83.3	89.8	86.9	95.2
.53-.62	99.8	97.7	95.8	90.7	96.3	92.5	97.1
.63-.72	100.0	100.0	99.0	96.3	99.1	99.1	99.0
.73-.82			99.0	97.2	100.0	100.0	100.0
.83-.92			100.0	97.2			

Note: The indexes in the middle set of rows of each subtable (−.22 to +.22) have values that are (approximately) not significantly different from zero—or chance targeting.

Table 7-6

Cumulative Percentage Distributions by School and Grade: *Phi* Coefficients Between CE Participation and Achievement Scores Dichotomized at the 35th Percentile

Reading

Index Value	School (N = 205)	Grade 1 (N = 143)	Grade 2 (N = 164)	Grade 3 (N = 171)	Grade 4 (N = 163)	Grade 5 (N = 153)	Grade 6 (N = 144)
−.62 - −.53	0.0	0.0	0.0	0.0	0.0	0.0	0.0
−.52 - −.43	0.0	0.0	0.0	0.0	0.0	0.0	0.0
−.42 - −.33	0.0	1.4	0.0	0.0	0.0	0.0	0.0
−.32 - −.23	0.0	4.2	0.0	0.0	0.0	0.0	0.0
−.22 - −.13	0.0	11.2	1.2	0.6	0.6	0.0	1.4
−.12 - −.03	1.0	24.5	3.6	2.9	3.1	1.3	3.5
−.02-.02	1.0	33.6	5.5	4.1	4.9	2.6	5.6
.03-.12	7.3	58.0	12.7	7.0	8.6	5.2	10.4
.13-.22	23.9	76.2	23.6	18.1	20.4	12.4	22.9
.23-.32	48.8	90.9	31.5	30.4	37.0	26.1	43.8
.33-.42	79.5	95.1	52.1	46.2	54.3	46.4	61.8
.43-.52	94.6	98.6	72.7	63.7	71.6	67.3	76.4
.53-.62	98.5	99.3	84.2	82.5	83.3	83.0	91.7
.63-.72	99.5	99.3	93.9	91.2	95.1	93.5	96.5
.73-.82	100.0	99.3	98.8	94.2	96.9	96.7	98.6
.83-.92		100.0	100.0	97.7	98.2	97.4	98.6

Math

Index Value	School (N = 161)	Grade 1 (N = 87)	Grade 2 (N = 98)	Grade 3 (N = 108)	Grade 4 (N = 108)	Grade 5 (N = 109)	Grade 6 (N = 105)
−.62 - −.53	0.0	0.0	0.0	0.0	0.0	0.0	0.0
−.52 - −.43	0.0	0.0	0.0	0.0	0.0	0.0	0.0
−.42 - −.33	0.0	1.1	0.0	0.9	0.0	0.0	0.0
−.32 - −.23	0.0	4.5	0.0	2.8	0.0	0.9	0.0
−.22 - −.13	1.2	10.1	2.0	3.7	2.8	0.9	1.9
−.12 - −.03	5.6	18.0	11.2	9.3	9.3	2.8	4.8
−.02-.02	8.1	31.5	15.3	13.0	10.2	6.4	10.5
.03-.12	26.1	48.3	29.6	23.1	25.0	12.8	20.0
.13-.22	53.4	66.3	49.0	35.2	41.7	33.0	51.4
.23-.32	79.5	83.1	63.3	52.8	68.5	51.4	68.6
.33-.42	93.2	92.9	80.6	73.1	82.4	74.3	81.9
.43-.52	98.1	95.5	90.8	84.3	92.6	88.1	91.4
.53-.62	98.8	96.6	95.9	91.7	99.1	92.7	98.1
.63-.72	100.0	98.9	100.0	96.3	99.1	98.2	98.1
.73-.82		98.9		97.2	100.0	100.0	100.0
.83-.92		98.9		97.2			

Note: The indexes in the middle set of rows of each subtable (−.22 to +.22) have values that are (approximately) not significantly different from zero—or chance targeting.

For each *phi* coefficient it is possible to determine the maximum value that it could take given the existing marginal frequencies. When the marginals are equal for the two variables, the maximum possible value for *phi* is 1.0. In general, especially when the marginals aren't equal, a theoretical maximum can be computed. Guilford (1965, p. 336 ff.) describes this maximum as phi_{max}, and points out that dividing *phi* by phi_{max} serves in part to correct the attentuated coefficient. The resulting coefficient is no longer a true correlation, and its distributional characteristics are not all desirable, but despite theoretical shortcomings it is of interest in the case at hand to impose this 'correction.' We can then determine whether an adjustment of this kind makes any important difference, and so draw some rough inferences about the impact of existing funding constraints (that would cause unequal marginals, in part) on targeting.

Table 7-7

Cumulative Percentage Distributions by School: *Phi* Coefficients Between CE Participation and Achievement, With Achievement Dichotomized at Three Different Percentile Points for Reading and Math

Index Value*	Reading			Math		
	50th Percentile	40th Percentile	35th Percentile	50th Percentile	40th Percentile	35th Percentile
-.20	0.0	0.0	0.0	0.6	0.0	0.0
-.15	0.0	0.0	0.0	1.2	0.6	1.2
-.10	0.0	1.0	0.5	3.1	1.9	1.9
-.05	1.5	1.5	1.0	6.8	4.3	5.6
.00	2.0	1.5	1.0	9.9	7.5	8.1
.05	3.9	2.9	3.4	16.1	15.5	13.7
.10	7.3	6.3	7.3	29.3	27.3	26.1
.15	15.6	13.7	13.2	43.5	41.0	42.2
.20	26.8	20.5	23.9	54.7	55.3	53.4
.25	37.1	33.2	35.6	73.9	73.9	68.3
.30	50.2	46.8	48.8	84.5	82.6	79.5
.35	62.0	60.0	63.9	90.1	90.7	89.4
.40	78.5	76.6	79.5	93.2	93.8	93.2
.45	88.3	86.3	86.3	95.7	95.7	96.3
.50	95.1	94.6	94.6	96.9	98.1	98.1
.55	98.0	98.0	97.1	98.1	98.1	98.8
.60	99.0	99.5	98.5	98.8	99.8	98.8
.65	99.5	99.5	99.5	98.8	99.4	99.4
.70	99.5	100.0	99.5	99.4	100.0	100.0
.75	99.5		100.0	100.0		
.80	100.0					

*Index values reported as midpoints of intervals of .05.

Table 7-8

Cumulative Percentage Distributions by School and Grade: *Phi* Coefficients Between CE Participation and Teacher Judgment of Need for CE, for Reading and Math

Reading

Index Value	School (N = 205)	Grade 1 (N = 138)	Grade 2 (N = 169)	Grade 3 (N = 172)	Grade 4 (N = 163)	Grade 5 (N = 153)	Grade 6 (N = 143)
−.42 - −.33	0.0	0.0	0.0	0.0	0.0	0.0	0.0
−.32 - −.23	0.0	1.4	0.0	0.0	0.0	0.0	0.0
−.22 - −.13	0.0	5.8	0.0	0.6	0.6	0.0	0.0
−.12 - −.03	2.0	10.9	1.8	2.3	2.5	1.3	2.1
−.02-.02	2.4	13.0	3.0	2.9	3.1	2.0	5.6
.03-.12	4.9	15.9	4.7	4.1	3.7	2.6	8.4
.13-.22	8.3	22.5	8.9	8.1	7.4	5.9	11.9
.23-.32	16.1	28.3	16.0	14.5	14.1	9.8	23.1
.33-.42	30.2	38.4	20.7	23.3	28.8	22.9	34.3
.43-.52	44.9	50.0	32.0	37.2	46.0	32.7	46.9
.53-.62	65.9	67.4	46.2	47.1	57.7	43.1	62.2
.63-.72	84.4	76.1	60.4	62.2	70.6	66.0	74.8
.73-.82	96.1	87.0	81.1	80.8	84.0	84.3	86.7
.83-.92	100.0	94.9	95.3	92.4	91.4	95.4	96.5

Math

Index Value	School (N = 161)	Grade 1 (N = 85)	Grade 2 (N = 95)	Grade 3 (N = 107)	Grade 4 (N = 108)	Grade 5 (N = 107)	Grade 6 (N = 102)
−.42 - −.33	0.0	1.2	0.0	0.0	0.0	0.0	0.0
−.32 - −.23	0.0	4.7	0.0	0.9	0.9	0.0	0.0
−.22 - −.13	0.6	8.2	1.1	3.7	2.8	0.0	1.0
−.12 - −.03	3.7	16.5	6.3	8.4	6.5	6.5	3.9
−.02-.02	5.6	17.6	11.6	9.3	8.3	8.4	4.9
.03-.12	9.9	28.2	14.7	14.0	12.0	12.1	6.9
.13-.22	21.7	38.8	24.2	19.6	20.4	19.6	15.7
.23-.32	39.1	45.9	33.7	29.9	31.5	30.8	32.4
.33-.42	58.4	57.6	48.4	43.9	46.3	43.0	44.1
.43-.52	73.9	69.4	67.4	63.6	62.0	54.2	53.9
.53-.62	84.5	75.3	76.8	74.8	74.1	67.3	74.5
.63-.72	94.4	87.1	83.2	81.3	84.3	80.4	89.2
.73-.82	98.1	95.3	90.5	92.5	92.6	86.9	93.1
.83-.92	99.4	97.6	95.8	95.3	97.2	93.5	95.1

The next group of tables are organized and formatted like those of the preceding sections. Table 7-9 contains the frequency distributions for schools and for each grade, based on the 40th percentile criterion for reading and for math. Table 7-10 contains the reading and math frequency distributions based on the 35th percentile. (The correlations derived from the 50th percentile tended to be lower in the preceding section, and this ordering would not likely be changed by the correction procedure. The 50th percentile correlations were therefore excluded from the adjusted-*phi* analyses.)

Table 7-9

Cumulative Percentage Distributions by School and Grade: Corrected *Phi* Coefficients Between CE Participation and Achievement Scores Dichotomized at the 40th Percentile

Reading

Index Value	School (N = 205)	Grade 1 (N = 143)	Grade 2 (N = 167)	Grade 3 (N = 171)	Grade 4 (N = 162)	Grade 5 (N = 154)	Grade 6 (N = 144)
Less than −.63	0.0	4.9	0.0	1.2	0.6	0.0	0.0
−.62 - −.53	0.0	6.3	0.0	1.2	0.6	0.6	0.0
−.52 - −.43	0.0	8.4	0.6	1.2	0.6	0.6	0.0
−.42 - −.33	0.5	9.1	1.2	1.2	0.6	0.6	0.7
−.32 - −.23	0.5	11.9	1.8	1.2	0.6	0.6	0.7
−.22 - −.13	0.5	16.8	2.4	1.8	1.2	1.3	0.7
−.12 - −.03	1.0	22.4	4.2	1.8	3.1	2.6	2.1
−.02-.02	1.0	30.1	6.6	2.9	4.3	3.9	4.2
.03-.12	2.0	38.5	9.0	4.7	4.9	3.9	4.9
.13-.22	4.4	50.3	11.4	5.3	6.2	5.2	9.0
.23-.32	10.2	63.6	15.0	7.6	12.3	8.4	13.2
.33-.42	22.4	70.6	19.8	12.3	17.9	12.3	16.0
.43-.52	32.7	79.7	27.5	17.0	24.1	19.5	22.9
.53-.62	52.2	88.1	38.9	30.4	38.9	29.9	35.4
.63-.72	73.2	93.0	53.9	46.8	51.2	44.8	47.9
.73-.82	85.9	94.4	64.7	59.6	60.5	58.4	59.0
.83-.92	95.6	95.1	72.5	64.3	64.2	63.6	63.9

Math

Index Value	School (N = 161)	Grade 1 (N = 86)	Grade 2 (N = 96)	Grade 3 (N = 108)	Grade 4 (N = 108)	Grade 5 (N = 107)	Grade 6 (N = 105)
Less than −.63	0.0	3.5	0.0	2.8	0.0	0.0	0.0
−.62 - −.53	0.0	8.1	0.0	3.7	0.0	0.0	0.0
−.52 - −.43	1.2	9.3	0.0	3.7	0.9	0.0	1.9
−.42 - −.33	1.2	14.0	1.0	3.7	1.9	0.0	1.9
−.32 - −.23	1.9	17.4	4.2	4.6	1.9	0.0	2.9
−.22 - −.13	1.9	22.1	6.3	5.6	2.8	2.8	3.8
−.12 - −.03	4.3	25.6	8.3	8.3	7.4	4.7	6.7
−.02-.02	6.8	29.1	13.5	10.2	8.3	4.7	8.6
.03-.12	11.8	32.6	19.8	13.0	15.7	5.6	12.4
.13-.22	15.5	45.3	20.8	16.7	19.4	10.3	20.0
.23-.32	26.1	53.5	32.3	21.3	31.5	14.0	27.6
.33-.42	39.8	61.6	42.7	27.8	39.8	22.4	33.3
.43-.52	54.7	67.4	51.0	36.1	52.8	37.4	44.8
.53-.62	67.1	72.1	58.3	49.1	62.0	43.9	52.4
.63-.72	76.4	76.7	65.6	55.6	68.5	55.1	64.8
.73-.82	84.5	80.2	70.8	61.1	75.0	61.7	66.7
.83-.92	88.8	81.4	70.8	64.8	75.0	64.5	68.6

Note: The indexes in the middle set of rows of each subtable (−.22 to +.22) have values that are (approximately) not significantly different from zero—or chance targeting.

Table 7-10

Cumulative Percentage Distributions by School and Grade: Corrected *Phi* Coefficients Between CE Participation and Achievement Scores Dichotomized at the 35th Percentile

Reading

Index Value	School (N = 205)	Grade 1 (N = 143)	Grade 2 (N = 165)	Grade 3 (N = 171)	Grade 4 (N = 162)	Grade 5 (N = 153)	Grade 6 (N = 144)
Less than −.63	0.0	2.1	0.0	1.2	0.6	0.0	0.0
−.62 - −.53	0.0	2.8	0.0	1.2	1.2	0.0	0.0
−.52 - −.43	0.0	4.9	0.6	1.2	1.2	0.0	0.0
−.42 - −.33	0.0	8.4	0.6	1.2	1.2	0.0	0.7
−.32 - −.23	0.0	12.6	1.8	1.2	1.2	0.0	0.7
−.22 - −.13	0.5	16.8	3.0	2.3	1.2	0.7	1.4
−.12 - −.03	1.0	29.4	3.6	2.9	3.1	1.3	3.5
−.02-.02	1.0	32.9	4.2	4.1	4.9	2.0	5.6
.03-.12	1.0	46.2	6.7	5.8	6.8	3.9	7.6
.13-.22	7.3	62.2	11.5	7.0	10.5	3.9	10.4
.23-.32	15.1	70.6	14.5	12.3	17.3	7.8	13.2
.33-.42	27.8	81.1	24.8	16.4	23.5	11.8	27.1
.43-.52	49.3	85.3	32.1	28.1	35.2	24.8	33.3
.53-.62	70.7	92.3	46.1	40.4	50.6	41.8	44.4
.63-.72	82.0	95.1	57.6	59.6	63.0	58.2	57.6
.73-.82	92.7	95.8	69.1	69.0	74.1	67.3	69.4
.83-.92	98.5	95.8	80.0	74.9	76.5	70.6	71.5

Math

Index Value	School (N = 161)	Grade 1 (N = 86)	Grade 2 (N = 96)	Grade 3 (N = 108)	Grade 4 (N = 108)	Grade 5 (N = 107)	Grade 6 (N = 105)
Less than −.63	0.6	5.8	0.0	2.8	1.9	0.0	0.0
−.62 - −.53	1.2	5.8	0.0	3.7	1.9	0.0	0.0
−.52 - −.43	1.2	8.1	0.0	3.7	1.9	0.0	1.9
−.42 - −.33	1.2	9.3	1.0	3.7	1.9	0.0	3.8
−.32 - −.23	1.9	11.6	4.2	5.6	2.8	0.0	3.8
−.22 - −.13	3.7	17.4	7.3	6.5	5.6	0.9	4.8
−.12 - −.03	5.0	22.1	10.4	9.3	10.2	2.8	5.7
−.02-.02	8.1	25.6	11.5	11.1	10.2	4.7	8.6
.03-.12	10.6	36.0	20.8	13.0	15.7	5.6	14.3
.13-.22	18.0	43.0	24.0	17.6	18.5	9.3	19.0
.23-.32	33.5	53.5	31.3	24.1	35.2	15.9	32.4
.33-.42	42.9	60.5	44.8	28.7	42.6	25.2	38.1
.43-.52	58.4	67.4	53.1	41.7	54.6	34.6	43.8
.53-.62	72.7	73.3	61.5	51.9	68.5	46.7	58.1
.63-.72	78.3	79.1	69.8	56.5	74.1	59.8	70.5
.73-.82	85.7	100.0	75.0	62.0	80.6	68.2	76.2
.83-.92	89.4		75.0	64.8	80.6	69.2	78.1

Note: The indexes in the middle set of rows of each subtable (−.22 to +.22) have values that are (approximately) not significantly different from zero—or chance targeting.

Applied to our data, the adjustment produces a small percentage of index values that are less than −1.0, i.e., less than the theoretical minimum for *phi* coefficients. The inappropriateness of the correction for negative values is one of the major deficiencies in the correction procedure. Most of the values are positive, however, and the correction clearly increases the *phi* values. For example, in the school-level analysis for the 40th percentile cutoff, the median value for the adjusted *phi* is about .59, compared to about .32 for the unadjusted *phi*. As a description of targeting, the adjusted *phi* is successful in eliminating the unwanted contributions of the marginal frequencies, and so is better for our purposes than the unadjusted *phi*. At the same time, the adjusted *phi* provides a picture of better targeting of CE.

Adjusted Phi Coefficients with Teacher Judgment of Need for CE. The adjustment was also applied to the calculation of *phi* coefficients when teacher judgment served as the definition of need for CE. The distributions of the resulting indexes appear in Table 7-11 for reading and math. The adjustment results in larger index values from those reported in Table 7-8 where the coefficients were not adjusted.

Point-Biserial Coefficients with Measures of Instructional Services Received. Having examined the relationship between the need indicators and CE status, we now turn our attention to the relationship between these same need indicators and measures of instructional services received by the students. To the extent that the intention of CE is that participants receive more or less costly services, targeting indexes based on services and need could be expected to better reflect the spirit of CE programs. An additional advantage of using a measure of services received is that even schools without formal CE programs could support the calculation of a targeting index (which indicates how well services are focused on the needy). We believe that this is a definite advantage because, even without formal CE programs, schools may do good jobs at targeting supplemental services to students in need of them.

Two measures of services are used: total exposure (i.e., hours of instruction attended during the school year) to reading and math instructions, and total resource cost for reading and math instruction. The former measure was taken from the Student Participation and Attendance Records for reading and for math (SPARM and SPAM, respectively), which were completed four times during the school year for four selected weeks. The latter is a cost-weighted composite of many kinds of instructional services and materials, and provides a way of quantifying a diverse set of services.

The point-biserial correlations for reading and math are presented in Table 7-12 for the CTBS scores, dichotomized at the 40th percentile. The corresponding correlations for the 35th percentile dichotomization appear in Table 7-13. The index values are not very difficult for the two criterion levels, although the 35th percentile cutoff yields slightly higher index values for reading, while the 40th is better for math. Again, the correlations are higher for reading than for math, but neither is very high. (The coefficients are lower than most of the *phi* coefficients based on selection status, but many non-CE schools have been included in the tabulation of the point-biserials. We can conclude that non-CE schools do not target services to low-achieving students as well as CE schools do.) The best

Table 7-11

Cumulative Percentage Distributions by School and Grade: Adjusted *Phi* Coefficients Between CE Participation and Teacher Judgment of Need for CE, for Reading and Math

Index Value	School (N = 205)	Grade 1 (N = 138)	Grade 2 (N = 169)	Grade 3 (N = 172)	Grade 4 (N = 163)	Grade 5 (N = 153)	Grade 6 (N = 143)
Reading							
Less than −.33	0.0	3.6	0.6	0.0	1.2	0.0	0.0
−.32 - −.23	0.5	5.1	0.6	0.6	1.2	0.7	0.0
−.22 - −.13	1.0	8.0	1.8	1.2	0.0	0.7	0.7
−.12 - −.03	2.0	10.9	1.8	2.3	2.5	1.3	2.8
−.02-.02	2.0	11.6	3.0	2.9	3.1	2.0	4.2
.03-.12	3.9	14.5	3.0	2.9	3.1	3.3	6.3
.13-.22	5.9	14.5	4.7	3.5	3.7	3.3	8.4
.23-.32	7.3	15.2	6.5	5.2	4.3	4.6	10.5
.33-.42	9.3	19.6	8.3	8.7	5.5	5.9	13.3
.43-.52	13.7	24.6	11.8	11.6	9.2	11.1	16.8
.53-.62	20.5	30.4	17.2	16.9	14.7	15.7	21.7
.63-.72	35.6	39.1	24.9	25.0	30.7	21.6	30.8
.73-.82	49.3	44.9	33.7	38.4	39.9	31.4	42.6
.83-.92	77.1	61.6	52.1	50.0	52.1	49.0	55.2
Math							
Index Value	**School (N = 161)**	**Grade 1 (N = 85)**	**Grade 2 (N = 95)**	**Grade 3 (N = 107)**	**Grade 4 (N = 108)**	**Grade 5 (N = 107)**	**Grade 6 (N = 102)**
Less than −.33	0.6	5.9	1.1	0.9	1.9	0.9	2.0
−.32 - −.23	1.9	7.1	2.1	1.9	3.7	1.9	2.9
−.22 - −.13	3.1	11.8	5.3	5.6	5.6	1.9	2.9
−.12 - −.03	5.0	16.5	8.4	8.4	6.5	6.5	3.9
−.02-.02	5.6	17.6	10.5	9.3	7.4	8.4	3.9
.03-.12	7.5	21.2	11.6	12.1	9.3	11.2	3.9
.13-.22	9.9	24.7	15.8	12.1	13.0	13.1	7.8
.23-.32	15.5	29.4	16.8	14.0	14.8	15.9	13.7
.33-.42	19.9	36.5	20.0	20.6	16.7	20.6	15.7
.43-.52	26.1	41.2	26.3	30.8	23.1	23.4	21.6
.53-.62	38.5	45.9	34.7	36.4	28.7	28.0	28.4
.63-.72	52.8	50.6	47.4	47.7	43.5	34.6	42.2
.73-.82	69.6	60.0	60.0	57.0	56.5	49.5	52.0
.83-.92	83.9	74.1	71.6	63.6	64.8	61.7	60.8

Table 7-12

Cumulative Percentage Distributions by School and Grade: Point-Biserial Correlations Between Total Exposure to Instruction and Achievement Scores Dichotomized at the 40th Percentile

Index Value	School (N = 206)	Grade 1 (N = 182)	Grade 2 (N = 181)	Grade 3 (N = 182)	Grade 4 (N = 177)	Grade 5 (N = 175)	Grade 6 (N = 162)
Reading							
Less than −.63	0.0	0.5	0.6	0.5	1.7	0.6	1.2
−.62 - −.53	0.0	1.1	1.7	0.5	2.3	0.6	2.5
−.52 - −.43	0.0	2.7	3.3	0.5	2.8	2.9	3.7
−.42 - −.33	0.0	6.6	8.3	4.9	6.2	4.6	6.8
−.32 - −.23	1.9	15.9	16.0	11.5	10.2	8.0	11.7
−.22 - −.13	7.8	28.0	24.9	19.2	16.4	13.1	22.2
−.12 - −.03	21.8	45.6	34.3	30.8	27.1	25.1	37.0
−.02-.02	35.0	59.3	45.3	37.4	32.2	32.6	43.8
.03-.12	69.9	76.4	60.2	48.9	49.7	43.4	54.9
.13-.22	89.3	90.1	74.6	64.8	63.3	60.6	69.1
.23-.32	95.1	94.0	81.2	76.4	71.8	74.9	79.6
.33-.42	99.0	96.7	88.4	83.0	79.7	82.9	85.8
.43-.52	99.5	98.9	90.1	91.2	87.0	91.4	93.8
.53-.62	100.0	99.5	92.8	96.7	93.2	96.0	97.5
.63-.72		99.5	97.8	98.4	98.3	97.1	99.4
.73-.82		99.5	99.4	98.9	99.4	98.9	100.0
.83-.92		100.0	100.0	98.9	100.0	99.4	

Index Value	School (N = 206)	Grade 1 (N = 182)	Grade 2 (N = 180)	Grade 3 (N = 179)	Grade 4 (N = 176)	Grade 5 (N = 174)	Grade 6 (N = 162)
Math							
Less than −.63	0.0	0.0	1.1	0.0	0.6	0.0	1.2
−.62 - −.53	0.0	0.5	2.8	1.7	1.7	1.1	1.9
−.52 - −.43	0.0	3.8	4.4	4.5	4.5	2.9	3.7
−.42 - −.33	0.5	6.0	9.4	5.6	8.0	8.6	6.8
−.32 - −.23	3.9	12.6	17.8	15.1	15.9	17.8	16.7
−.22 - −.13	17.0	23.6	27.2	26.3	28.4	31.0	32.1
−.12 - −.03	47.6	51.6	48.3	44.7	46.6	46.6	53.7
−.02-.02	65.0	59.3	58.9	50.8	55.7	59.2	61.7
.03-.12	89.3	75.8	73.9	72.1	71.0	71.8	75.9
.13-.22	96.1	86.8	89.4	82.7	83.5	79.3	86.4
.23-.32	99.0	92.9	95.6	91.1	89.8	87.9	93.8
.33-.42	99.5	97.8	98.3	95.5	96.0	93.1	97.5
.43-.52	100.0	99.5	99.4	97.2	97.2	96.0	98.1
.53-.62		100.0	99.4	99.4	98.3	98.9	98.8
.63-.72			100.0	99.4	99.4	99.4	99.4
.73-.82				99.4	100.0	99.4	100.0
.83-.92				99.4		100.0	

Note The indexes in the middle set of rows of each subtable (−.22 to +.22) have values that are (approximately) not significantly different from zero—or chance targeting of services.

Table 7-13

Cumulative Percentage Distributions by School and Grade: Point-Biserial Correlations Between Total Exposure to Instruction and Achievement Scores Dichotomized at the 35th Percentile

Reading

Index Value	School (N = 206)	Grade 1 (N = 182)	Grade 2 (N = 179)	Grade 3 (N = 180)	Grade 4 (N = 176)	Grade 5 (N = 174)	Grade 6 (N = 162)
Less than −.63	0.0	0.5	1.7	0.6	1.1	0.6	1.2
−.62 - −.53	0.0	1.1	2.8	1.1	1.7	0.6	3.1
−.52 - −.43	0.0	2.7	4.5	2.2	2.3	3.4	3.7
−.42 - −.33	0.0	8.2	7.3	3.9	5.7	5.2	6.2
−.32 - −.23	1.9	14.8	12.8	11.7	10.8	8.6	13.0
−.22 - −.13	7.8	31.3	25.1	20.0	15.3	16.1	22.8
−.12 - −.03	24.8	50.0	36.3	29.4	28.4	25.9	35.2
−.02-.02	38.8	57.7	43.6	35.6	34.7	32.8	45.7
.03-.12	73.8	77.5	60.3	51.7	50.6	44.3	56.8
.13-.22	91.3	86.8	75.4	65.0	60.8	62.1	69.1
.23-.32	94.7	93.4	85.5	75.6	70.5	73.6	80.2
.33-.42	99.0	96.2	88.3	82.2	79.0	83.3	87.7
.43-.52	100.0	99.5	92.2	90.0	84.1	90.8	92.6
.53-.62		99.5	96.6	96.7	92.6	94.3	98.8
.63-.72		99.5	98.3	98.3	95.5	96.6	99.4
.73-.82		100.0	99.4	98.9	98.9	97.7	99.4
.83-.92			100.0	98.9	100.0	99.4	99.4

Math

Index Value	School (N = 206)	Grade 1 (N = 182)	Grade 2 (N = 180)	Grade 3 (N = 179)	Grade 4 (N = 176)	Grade 5 (N = 174)	Grade 6 (N = 162)
Less than −.63	0.0	0.0	1.1	0.0	0.6	0.0	0.6
−.62 - −.53	0.0	0.0	1.7	0.6	1.1	0.6	1.2
−.52 - −.43	0.0	2.7	5.0	3.4	3.4	4.0	2.5
−.42 - −.33	1.0	4.4	9.4	6.7	7.4	8.6	4.3
−.32 - −.23	1.5	9.9	16.7	13.4	13.6	17.8	12.3
−.22 - −.13	15.0	23.1	29.4	23.5	28.4	31.0	29.0
−.12 - −.03	46.1	46.7	46.7	46.9	46.6	50.0	54.3
−.02-.02	58.7	55.5	56.7	55.9	56.3	56.3	64.2
.03-.12	86.9	74.7	76.1	70.9	73.9	73.6	79.0
.13-.22	95.1	86.8	91.7	84.9	85.8	79.3	88.3
.23-.32	99.5	92.3	95.0	90.5	92.0	88.5	95.1
.33-.42	99.5	97.3	98.3	93.9	96.6	92.5	98.1
.43-.52	100.0	98.9	98.9	97.2	98.3	96.0	98.1
.53-.62		100.0	99.4	99.4	98.9	98.9	98.8
.63-.72			100.0	99.4	100.0	100.0	99.4
.73-.82				99.4			100.0
.83-.92				99.4			

cutoff for reading (35th percentile for schools) yields 38.8 percent non-positive index values, and for math (40th percentile for schools) it yields 65.0 percent non-positive index values Table 7-14 contains the cumulative distributions for reading and math of the indexes created by correlating teacher judgment of student need for CE with exposure to instruction (reading need with reading instruction, math need with math instruction). The percentage of non-positive correlations is 34.0 for reading and 56.8 for math—not much better than for the previous criterion.

Finally, the same set of CTBS dichotomous scores and the teacher judgment variable are correlated with the corresponding resource cost measures in Tables 7-15 to 7-17. The correlations based on resource cost are larger than those based on services. Again, there is little difference between the correlations obtained for the two percentile cutoffs, and again the correlations for reading are considerably higher than those for math. Teacher judgment is a better correlate of level of service than either achievement-score dichotomy, just as it is a better correlate of CE selection (see previous remarks on limits to the validity of indexes using this measure).

INDEXES BASED ON CORRELATIONS WITH CTBS PERCENTILE SCORES

In the previous sections, several indexes have been derived from correlations between dichotomized CTBS scores and CE status or measures of CE services received. It is also of interest to examine indexes that are based on correlations involving the CTBS scores themselves. The form of the achievement scores used for these analyses is the percentile, as we expected that the percentile scores would have more appropriate distributions within grades and within schools. The indexes to be described were created by correlating, for both reading and math, the percentile scores with CE status, with total exposure to instruction, and with resource cost of instruction. Because the CE-status variable is dichotomous, the correlation between it and the CTBS score is a point-biserial correlation. The variables reflecting the amount of instructional services received, however, are continuous and quantitative variables; thus the correlations with the CTBS scores are Pearson product-moment coefficients.

The correlations between the CTBS scores and CE status are the indexes in Table 7-18. The table contains the cumulative frequencies for the indexes by grade and by school. Because low percentile scores (negative when standardized) are associated with CE participation (positive when standardized) when good targeting occurs, good targeting would be indicated by negative correlation coefficients. In order to make the indexes of this section comparabale to those already studied, and in line with intuition, we have reflected all the coefficients (changed their signs) in this section.

Except for the first grade, the correlations for reading are overwhelmingly positive (i.e., indicating good targeting), with only .6 percent to 3.6 percent of the values negative. The median value at the school level is between .35 and .45, however, indicating only moderately effective targeting. As with the previous index, the math values indicate somewhat poorer targeting than those for reading.

Table 7-14

Cumulative Percentage Distributions by School and Grade: Point-Biserial Correlations Between Total Exposure to Instruction and Teacher Judgment of Student Need for CE

Reading

Index Value	School (N = 206)	Grade 1 (N = 175)	Grade 2 (N = 181)	Grade 3 (N = 184)	Grade 4 (N = 177)	Grade 5 (N = 173)	Grade 6 (N = 159)
Less than -.63	0.0	2.3	0.6	0.5	0.0	0.0	1.3
-.62 - -.53	0.0	2.3	0.6	0.5	0.6	0.0	1.3
-.52 - -.43	0.0	4.6	2.8	2.7	2.8	2.3	3.8
-.42 - -.33	0.5	13.1	8.3	4.9	5.1	5.8	5.0
-.32 - -.23	1.9	22.9	15.5	7.6	6.8	9.2	15.1
-.22 - -.13	9.7	33.1	23.8	15.2	14.7	15.0	21.4
-.12 - -.03	23.8	49.7	34.8	28.8	26.0	24.9	34.0
-.02-.02	34.0	56.0	40.3	32.1	28.8	30.6	41.5
.03-.12	63.6	72.6	53.6	40.8	41.2	42.8	51.6
.13-.22	83.0	82.9	65.2	54.3	50.3	54.3	58.5
.23-.32	91.3	88.6	73.5	65.2	61.0	64.2	66.0
.33-.42	96.6	92.6	79.0	75.5	69.5	71.1	74.8
.43-.52	97.1	96.0	83.4	81.0	76.8	79.2	81.8
.53-.62	98.5	96.0	90.6	88.0	80.2	85.5	86.8
.63-.72	100.0	97.7	94.5	92.9	85.3	90.2	90.6
.73-.82		98.3	98.3	95.7	91.0	93.6	95.0
.83-.92		100.0	100.0	97.8	96.6	97.7	99.4

Math

Index Value	School (N = 206)	Grade 1 (N = 173)	Grade 2 (N = 172)	Grade 3 (N = 173)	Grade 4 (N = 173)	Grade 5 (N = 171)	Grade 6 (N = 158)
Less than -.63	0.0	0.0	1.7	1.7	0.0	1.2	0.0
-.62 - -.53	0.0	1.2	2.3	2.9	1.7	2.3	2.5
-.52 - -.43	0.5	2.3	2.9	4.6	3.5	4.7	7.0
-.42 - -.33	1.5	11.0	6.4	8.7	11.6	9.9	10.1
-.32 - -.23	5.8	23.7	16.9	17.3	15.6	16.4	18.4
-.22 - -.13	18.4	38.2	32.0	28.3	26.0	32.2	31.0
-.12 - -.03	41.7	54.9	46.5	42.8	40.5	45.6	46.2
-.02-.02	56.8	61.8	51.7	52.0	47.4	54.4	57.0
.03-.12	73.3	76.9	68.0	67.1	61.3	64.9	68.4
.13-.22	87.4	86.7	79.7	78.0	76.3	77.2	76.6
.23-.32	95.1	91.9	84.9	85.0	80.9	82.5	87.3
.33-.42	97.1	95.4	88.4	89.0	86.1	85.4	90.5
.43-.52	98.5	97.1	91.9	94.2	89.6	90.6	91.8
.53-.62	99.5	99.4	94.8	96.5	96.0	94.2	94.9
.63-.72	100.0	99.4	97.7	98.3	97.7	96.5	96.2
.73-.82		100.0	99.4	98.8	98.8	96.5	98.1
.83-.92			99.4	99.4	99.4	98.8	98.7

Note: The indexes in the middle set of rows of each subtable (-.22 to +.22) have values that are (approximately) not significantly different from zero—or chance targeting of services.

134

Table 7-15

Cumulative Percentage Distributions by School and Grade: Point-Biserial Correlations Between Resource Cost of Instruction and Achievement Scores Dichotomized at the 40th Percentile

Reading

Index Value	School (N = 206)	Grade 1 (N = 182)	Grade 2 (N = 181)	Grade 3 (N = 180)	Grade 4 (N = 177)	Grade 5 (N = 175)	Grade 6 (N = 161)
Less than −.63	0.0	0.0	0.6	0.6	0.0	0.0	0.6
−.62 - −.53	0.0	0.0	1.7	0.6	0.0	0.0	0.6
−.52 - −.43	0.5	1.6	2.8	2.2	1.1	0.6	1.2
−.42 - −.33	0.5	5.5	5.0	3.3	1.1	2.3	3.7
−.32 - −.23	0.5	7.1	5.5	6.1	5.6	2.9	5.6
−.22 - −.13	2.4	17.0	7.2	8.8	7.9	5.2	9.9
−.12 - −.03	6.3	33.5	13.3	10.6	13.6	7.4	14.3
−.02-.02	10.2	42.9	15.5	11.7	16.9	12.0	18.0
.03-.12	24.3	58.8	23.2	21.0	20.9	17.7	26.1
.13-.22	50.0	76.4	35.4	28.9	30.5	29.7	40.4
.23-.32	81.1	85.7	43.1	39.4	42.4	46.3	58.4
.33-.42	94.2	92.9	56.4	52.2	56.5	64.0	72.7
.43-.52	98.5	97.8	73.5	66.7	69.5	77.1	85.1
.53-.62	99.5	99.5	89.5	83.9	86.4	89.1	93.2
.63-.72	100.0	99.5	95.0	93.3	97.7	94.9	98.1
.73-.82		99.5	98.9	97.8	98.9	97.7	99.4
.83-.92		100.0	100.0	99.4	98.9	98.9	100.0

Math

Index Value	School (N = 206)	Grade 1 (N = 182)	Grade 2 (N = 180)	Grade 3 (N = 178)	Grade 4 (N = 176)	Grade 5 (N = 174)	Grade 6 (N = 161)
Less than −.63	0.0	0.0	0.0	0.6	0.0	0.0	0.0
−.62 - −.53	0.0	0.0	0.6	1.7	0.0	1.1	0.6
−.52 - −.43	0.0	1.1	0.6	2.2	1.7	2.3	1.2
−.42 - −.33	0.5	1.6	5.0	4.5	5.7	4.6	3.1
−.32 - −.23	1.0	6.6	13.3	10.7	9.1	10.9	8.1
−.22 - −.13	4.9	13.2	18.9	20.8	12.5	16.7	17.4
−.12 - −.03	17.0	28.6	30.6	29.2	25.0	27.0	24.2
−.02-.02	32.5	39.6	37.8	33.7	35.2	35.6	30.4
.03-.12	55.3	58.2	52.2	46.6	49.4	48.9	47.2
.13-.22	81.1	76.4	71.7	59.0	67.6	61.5	63.4
.23-.32	95.6	89.0	81.1	77.5	78.4	71.8	77.0
.33-.42	98.1	94.5	91.1	88.8	88.6	82.2	90.1
.43-.52	99.5	98.9	96.7	92.1	94.3	89.7	95.7
.53-.62	100.0	99.5	98.3	95.5	97.7	97.1	98.8
.63-.72		100.0	99.4	98.3	100.0	99.4	100.0
.73-.82			100.0	98.9		100.0	
.83-.92				98.9			

Note: The indexes in the middle set of rows of each subtable (−.22 to +.22) have values that are (approximately) not significantly different from zero—or chance targeting of service costs.

Table 7-16

Cumulative Percentage Distributions by School and Grade: Point-Biserial Correlations Between Resource Cost of Instruction and Achievement Scores Dichotomized at the 35th Percentile

			Reading				
Index Value	**School (N = 206)**	**Grade 1 (N = 182)**	**Grade 2 (N = 179)**	**Grade 3 (N = 178)**	**Grade 4 (N = 176)**	**Grade 5 (N = 174)**	**Grade 6 (N = 161)**
Less than −.63	0.0	0.0	1.1	0.6	0.0	0.0	0.6
−.62 - −53	0.0	0.0	2.2	0.6	0.0	0.0	0.6
− 52 - −.43	0.0	0 5	2.8	1.7	0.0	0.6	0.6
− 42 - −.33	0.5	4 4	3.4	2.8	0.6	2.3	3.1
−.32 - −.23	1.5	10.4	3.9	5.1	2.8	2.3	5.6
− 22 - − 13	2 4	18 7	8.9	7 9	7.4	5.2	9.9
−.12 - − 03	6 3	36.8	13.4	10.1	14.2	8.0	13.7
−.02-.02	9 2	45 6	15 6	13.5	17 0	11.5	19.9
03-.12	24.8	66.5	22.3	19.7	21.6	16.7	27.3
13- 22	53.9	80 2	35 2	27.5	30 7	27.0	36.6
23-.32	81.6	89 0	46.4	39.3	42.6	40.8	57.8
33- 42	93.7	95.1	60.9	50.0	56.8	58.6	73.9
.43-.52	98.1	98.4	78.2	66.9	69.3	78.7	83.2
.53-.62	100.0	99.5	87 7	81.5	84.1	89.1	94.4
.63-.72		99.5	93.3	93.3	94.3	94.8	97.5
.73-.82		100 0	97.8	97.2	96.6	97.1	98.8
83-.92			99.4	99.4	98.3	98.9	99.4
			Math				
Index Value	**School (N = 206)**	**Grade 1 (N = 182)**	**Grade 2 (N = 180)**	**Grade 3 (N = 178)**	**Grade 4 (N = 176)**	**Grade 5 (N = 174)**	**Grade 6 (N = 161)**
Less than −.63	0.0	0.0	0.0	0.6	0.0	0.0	0.0
−.62 - −.53	0.0	0.0	0.6	1.1	0.0	0.6	0.6
−.52 - −.43	0.0	0.5	1.1	1.1	1.7	1.7	1.2
−.42 - −.33	0.5	1.6	5.0	3.4	5.1	3.4	4.3
−.32 - −.23	1.0	5.5	9.4	8.4	9.1	11.5	7.5
−.22 - −.13	4.9	13.2	20.6	18.5	14.8	18.4	15.5
−.12 - −.03	17.0	30.2	30.6	29.2	25.6	27.0	24.8
−.02-.02	30.1	40.7	39.4	33.7	37.5	35.6	28.6
.03-.12	52.4	61.5	51.7	46.1	51.7	47.1	46.6
.13-.22	79.6	75.3	67.8	60.1	67.6	61.5	64.0
.23-.32	95.1	84.1	81.7	78.1	80.1	71.8	78.3
.33-.42	98.1	91.2	90.6	86.5	88.6	81.6	87.6
.43-.52	99.0	97.8	94.4	92.1	96.6	89.1	95.0
.53-.62	100.0	98.9	98.9	96.1	98.3	97.1	98.8
.63-.72		99.5	99.4	97.8	99.4	100.0	99.4
.73-.82		99.5	100.0	98.9	100.0		100.0
83-.92		99.5		98.9			

Note The indexes in the middle set of rows in each subtable (−.22 to +.22) have values that are (approximately) not signficiantly different from zero—or chance targeting of service costs.

Table 7-17

Cumulative Percentage Distributions by School and Grade: Point-Biserial Correlations Between Resource Cost of Instruction and Teacher Judgment of Student Need for CE

Reading

Index Value	School (N = 206)	Grade 1 (N = 175)	Grade 2 (N = 181)	Grade 3 (N = 182)	Grade 4 (N = 177)	Grade 5 (N = 173)	Grade 6 (N = 158)
Less than −.63	0.0	0.6	1.1	1.6	0.0	0.0	1.9
−.62 - −.53	0.0	0.6	1.1	1.6	0.0	0.6	1.9
−.52 - −.43	0.0	2.9	2.2	2.2	0.6	1.2	1.9
−.42 - −.33	0.0	6.9	2.8	5.5	2.8	1.7	3.2
−.32 - −.23	0.0	8.6	5.0	5.5	4.0	4.0	5.7
−.22 - −.13	0.5	13.7	7.2	6.6	4.0	6.4	8.9
−.12 - −.03	3.4	22.9	12.2	9.3	6.8	8.7	12.0
−.02-.02	5.3	26.3	13.8	9.9	10.7	11.6	15.2
.03-.12	15.0	33.7	17.1	14.3	16.9	16.2	22.2
.13-.22	31.1	43.4	27.1	19.2	19.2	21.4	26.6
.23-.32	51.5	49.1	35.9	36.2	28.8	30.1	34.2
.33-.42	69.9	62.9	45.9	37.9	37.9	39.9	46.2
.43-.52	85.4	72.6	56.4	48.9	50.3	54.3	56.3
.53-.62	95.1	78.9	65.2	62.1	62.7	71.1	69.0
.63-.72	99.0	87.4	78.5	75.8	73.4	80.3	77.2
.73-.82	100.0	93.1	90.1	89.0	88.1	91.3	89.2
.83-.92		98.3	98.3	94.5	95.5	97.1	98.1

Math

Index Value	School (N = 206)	Grade 1 (N = 173)	Grade 2 (N = 172)	Grade 3 (N = 172)	Grade 4 (N = 173)	Grade 5 (N = 171)	Grade 6 (N = 158)
Less than −.63	0.0	0.0	0.6	0.0	0.6	1.8	0.6
−.62 - −.53	0.0	0.0	2.3	1.2	2.3	2.3	1.9
−.52 - −.43	0.0	0.6	2.3	2.9	4.0	5.3	3.2
−.42 - −.33	0.5	4.0	2.9	5.8	5.8	7.6	4.4
−.32 - −.23	1.5	8.7	8.7	8.1	9.2	14.0	6.3
−.22 - −.13	4.4	12.1	15.7	12.8	13.3	17.5	12.7
−.12 - −.03	9.7	22.5	24.4	20.3	20.2	25.1	22.8
−.02-.02	20.4	32.9	28.5	24.4	23.1	32.2	27.2
.03-.12	35.9	48.0	44.2	34.9	32.9	39.8	34.2
.13-.22	56.3	56.1	54.7	44.8	42.2	45.6	46.2
.23-.32	73.8	68.2	62.8	57.0	56.1	57.3	52.5
.33-.42	85.4	76.3	69.8	70.3	67.1	64.9	62.7
.43-.52	92.7	82.1	78.5	77.9	78.0	76.0	72.8
.53-.62	98.1	88.4	85.5	86.6	85.0	81.9	81.0
.63-.72	99.0	92.5	91.3	93.6	91.9	89.5	89.9
.73-.82	99.5	97.1	98.3	96.5	97.1	94.2	94.3
.83-.92	100.0	98.3	98.8	97.7	99.4	98.8	98.7

Note The indexes in the middle set of rows in each subtable (−.22 to +.22) have values that are (approximately) not significantly different from zero—or chance targeting of service costs.

Table 7-18

Cumulative Percentage Distributions by School and Grade: Point-Biserial Coefficents Between CE Participation and Achievement Percentile Scores

Reading

Index Value	School (N = 205)	Grade 1 (N = 143)	Grade 2 (N = 169)	Grade 3 (N = 172)	Grade 4 (N = 164)	Grade 5 (N = 156)	Grade 6 (N = 144)
Less than −.43	0.0	0.7	0.0	0.0	0.0	0.0	0.0
−.42 - −.33	0.0	2.1	0.0	0.0	0.0	0.0	0.0
−.32 - −.23	0.0	3.5	0.0	0.6	0.0	0.0	0.0
−.22 - −.13	0.0	8.4	1.2	1.7	0.0	0.6	0.0
−.12 - −.03	1.0	16.8	3.6	3.5	0.6	0.6	2.1
−.02-.02	2.0	24.5	4.1	4.7	1.8	1.3	2.1
.03-.12	5.9	46.2	9.5	5.8	7.3	3.2	7.6
.13-.22	20.0	69.9	17.8	13.4	18.3	9.0	22.9
.23-.32	41.5	83.9	27.2	25.0	31.1	19.9	39.6
.33-.42	66.3	94.4	41.4	37.8	47.6	40.4	59.7
.43-.52	91.2	97.9	54.4	57.6	70.1	63.5	73.6
.53-.62	98.5	98.6	74.0	80.2	84.1	79.5	88.2
.63-.72	99.5	98.6	91.7	91.3	93.3	94.9	95.8
.73-.82	100.0	99.3	98.8	98.3	97.0	96.2	97.9
.83-.92		99.3	99.4	98.3	99.4	98.7	98.6

Math

Index Value	School (N = 162)	Grade 1 (N = 86)	Grade 2 (N = 96)	Grade 3 (N = 109)	Grade 4 (N = 108)	Grade 5 (N = 108)	Grade 6 (N = 105)
Less than −.43	0.0	1.2	1.2	0.0	0.0	0.0	0.0
−.42 - −.33	0.0	2.3	1.0	4.6	0.9	0.0	0.0
−.32 - −.23	0.0	3.5	1.0	5.5	1.9	0.0	0.0
−.22 - −.13	1.9	11.6	3.1	7.3	3.7	0.0	1.0
−.12 - −.03	6.8	17.4	3.1	10.1	3.7	0.0	3.8
−.02-.02	10.5	25.6	8.3	11.0	7.4	0.9	4.8
.03-.12	24.7	39.5	24.0	16.5	13.0	11.1	14.3
.13-.22	45.3	59.3	42.7	25.7	36.1	23.1	40.0
.23-.32	77.8	83.7	55.2	51.4	53.7	42.6	65.7
.33-.42	91.2	88.4	77.1	68.8	80.6	68.5	80.0
.43-.52	95.7	93.0	90.6	81.7	89.8	81.5	87.6
.53-.62	98.8	98.8	96.9	89.0	96.3	90.7	96.2
.63-.72	99.4	100.0	97.9	95.4	98.1	98.1	99.0
.73-.82	100.0		99.0	96.3	100.0	99.1	100.0
.83-.92			100.0	97.2		100.0	

Note All coefficients have reversed signs, so high indexes indicate 'good' targeting. The indexes in the middle set of rows in each subtable (−.22 to +.22) have values that are (approximately) not significantly different from zero—or chance targeting.

The indexes described in Table 7-19 are the correlations between the same reading and math subtest scores and the total exposure to reading and math instruction. The positive correlations are again indicators of relatively good targeting. Although this measure of instructional service cannot be analyzed into CE and non-CE components, it is of interest to examine the amount of agreement between achievement level and amount of reading and math instruction to which the student is exposed, as an indicator of the degree to which need is associated with amount of services delivered to the student. The values of these indexes have surprisingly negative distributions. For reading, the median values for all grades and for the entire school are lower (i.e., worse) than .20; for math they are all lower than −.05. These median values indicate that there is precious little association between percentile scores (need) and numbers of hours of exposure to reading and math. Such a small overall effect may reflect the grossness of the service measure; every student in a school receives approximately the same number of hours, so hours may be an insensitive measure. Instead, perhaps, we should examine the quality of the services received.

The resource cost measures represent an attempt to get at the quality of the services by defining them according to their relative costs. For example, an hour of individual instruction with a special teacher is costed at a higher level than an hour in a classroom with a teaching assistant Table 7-20 presents the correlations between the CTBS scores and the resource-cost measures. The indexes show stronger (i.e., more positive) relationships than those based on hours of services received, demonstrating that targeting is going on in the manner expected (if it weren't, improving the variables would make the indexes lower). The medians for the indexes are more positive, about +.40 for reading for most grades and from about +.10 to +.20 for math. The values still do not indicate a strong relationship, however, with most correlations falling at or below the level required for positive statistical significance at the .05 level.

In summary, although it seemed worthwhile to examine indexes based on percentile scores, these indexes are only slightly more sensitive to targeting than the indexes considered previously. We should not be very surprised at this, because the dichotomies are merely coarse groupings of the percentiles, so the correlations based on them are expected to be smaller. Both percentiles and dichotomies are based on national norms, while good targeting of services to students ought to be based on a finer tuning within each school. Thus, neither the indexes of this section nor those of previous sections are finely tuned to the *relative* needs of students within a given school.

INDEXES BASED ON ACHIEVEMENT RANKS WITHIN SCHOOLS

Perhaps the most stringent concept of targeting is that, given a set of N students, of whom only n can be selected for compensatory services, the optimal allocation of services occurs when the n most needy of the set are selected. Each of the various subsettings of the n students into two groups (selected and not selected) can be given a value that reflects how well the subset focuses CE on the most needy *and* not on the least needy. A statistic suggested by Albert Beaton of the Educational Testing Service is based on this definition of optimal targeting. It assigns a value to any subset (of size n) of the entire population of size N, derived from the number of subsets of the same size that would be less targeted. For

Table 7-19

Cumulative Percentage Distributions by School and Grade: Product-Moment Coefficients Between Total Exposure to Instruction and Achievement Percentile Scores

Reading

Index Value	School (N = 206)	Grade 1 (N = 183)	Grade 2 (N = 183)	Grade 3 (N = 184)	Grade 4 (N = 180)	Grade 5 (N = 178)	Grade 6 (N = 162)
Less than −.43	0.0	1.1	4.4	1.1	3.3	1.7	3.7
−.42 - −.33	0.5	7.7	9.3	6.0	5.6	3.9	7.4
−.32 - −.23	3.4	17.5	18.6	14.1	9.4	8.4	11.7
−.22 - −.13	9.7	29.0	25.7	21.2	15.0	18.0	24.7
−.12 - −.03	27.7	45.9	37.7	29.3	28.3	29.2	37.7
−.02-.02	42.7	55.7	45.4	35.3	34.4	33.7	44.4
.03-.12	71.4	73.2	64.2	48.4	47.8	47.8	57.4
.13-.22	87.9	88.5	68.9	63.6	59.4	60.7	67.9
.23-.32	95.1	92.3	77.0	72.8	66.1	74.7	77.2
.33-.42	98.5	97.3	85.2	82.6	77.8	81.5	82.1
.43-.52	99.5	97.8	88.0	88.6	85.0	88.2	90.1
.53-.62	99.5	97.8	91.8	95.1	92.8	92.7	95.7
.63-.72	100.0	99.5	99.5	97.3	97.8	97.2	98.1
.73-.82		100.0	99.5	97.8	99.4	98.9	99.4
.83-.92			100.0	97.8	99.4	100.0	100.0

Math

Index Value	School (N = 206)	Grade 1 (N = 183)	Grade 2 (N = 183)	Grade 3 (N = 183)	Grade 4 (N = 180)	Grade 5 (N = 177)	Grade 6 (N = 162)
Less than −.43	0.0	3.3	6.6	5.5	4.4	3.4	3.7
−.42 - −.33	1.0	4.9	13.7	9.8	11.3	7.9	7.4
−.32 - −.23	5.8	13.7	20.2	15.3	17.8	16.9	18.5
−.22 - −.13	18.0	30.1	31.7	30.6	27.8	33.3	36.4
−.12 - −.03	46.6	55.2	47.0	47.0	48.9	50.3	59.3
−.02-.02	64.6	60.1	60.1	54.1	59.4	56.5	67.3
.03-.12	87.9	76.0	75.4	69.9	70.6	68.9	80.2
.13-.22	95.1	86.9	85.8	84.2	83.3	77.4	87.7
.23-.32	99.0	92.9	94.5	90.7	90.0	86.4	92.6
.33-.42	99.5	97.8	97.3	94.0	96.1	92.1	97.5
.43-.52	100.0	98.9	97.8	96.2	97.8	94.9	97.5
.53-.62		100.0	98.9	98.4	98.3	96.6	98.1
.63-.72			99.5	98.9	99.4	98.3	99.4
.73-.82			99.5	98.9	100.0	99.4	100.0
.83-.92			100.0	98.9		99.4	

Note: All coefficients have reversed signs, so high indexes indicate 'good' targeting. The indexes in the middle set of rows in each subtable (−.22 to +.22) have values that are (approximately) not significantly different from zero—or chance targeting of services.

Table 7-20

Cumulative Percentage Distributions by School and Grade: Product-Moment Coefficients Between Resource Cost of Instruction and Achievement Percentile Scores

Reading

Index Value	School (N = 206)	Grade 1 (N = 183)	Grade 2 (N = 183)	Grade 3 (N = 182)	Grade 4 (N = 180)	Grade 5 (N = 178)	Grade 6 (N = 161)
Less than −.43	0.5	2.7	2.7	3.8	0.6	1.7	1.9
−.42 - −.33	0.5	4.9	3.3	6.6	1.1	2.8	4.3
−.32 - −.23	1.0	7.7	4.4	7.7	3.9	3.4	5.6
−.22 - −.13	1.9	19.7	6.6	8.8	7.2	7.3	8.7
−.12 - −.03	8.3	34.4	11.5	12.1	11.7	11.2	13.7
−.02-.02	10.7	42.1	14.2	15.9	14.4	11.8	16.8
.03-.12	24.3	59.0	22.4	19.2	22.8	17.4	26.1
.13-.22	51.0	74.3	30.1	29.7	27.8	25.8	33.5
.23-.32	78.2	86.9	38.8	37.9	37.2	36.5	49.7
.33-.42	92.2	94.5	51.4	52.2	51.1	59.0	66.5
.43-.52	98.5	97.8	65.6	66.5	67.2	73.0	79.5
.53-.62	99.0	99.5	82.0	81.3	85.6	83.1	90.7
.63-.72	100.0	100.0	93.4	92.3	96.1	92.1	95.0
.73-.82			98.9	97.8	99.4	97.8	98.8
.83-.92			100.0	98.4	99.4	99.4	100.0

Math

Index Value	School (N = 206)	Grade 1 (N = 183)	Grade 2 (N = 183)	Grade 3 (N = 181)	Grade 4 (N = 180)	Grade 5 (N = 177)	Grade 6 (N = 161)
Less than −.43	0.0	1.6	2.3	4.4	1.7	1.7	1.9
−.42 - −.33	0.5	4.4	8.7	8.3	7.2	5.6	5.6
−.32 - −.23	1.0	9.3	16.4	13.8	10.0	11.3	9.3
−.22 - −.13	7.3	16.9	21.3	21.0	17.2	18.1	14.3
−.12 - −.03	18.9	31.7	28.4	27.1	23.9	29.4	23.6
−.02-.02	29.6	41.0	35.5	34.3	29.4	33.9	27.0
.03-.12	50.0	57.9	48.6	47.0	43.9	45.2	41.6
.13-.22	73.8	72.7	63.9	59.7	60.6	55.4	55.3
.23-.32	94.7	83.1	77.0	71.8	78.3	66.7	72.7
.33-.42	97.4	92.3	88.5	84.5	85.0	76.8	85.7
.43-.52	99.5	96.7	94.5	90.1	92.8	86.4	94.4
.53-.62	100.0	99.5	97.8	94.5	98.3	94.9	97.5
.63-.72		100.0	98.9	96.7	100.0	98.3	100.0
.73-.82			98.9	98.3		99.4	
.83-.92			100.0	98.3		100.0	

Note: All coefficients have reversed signs, so high indexes indicate 'good' targeting. The indexes in the middle set of rows in each subtable (−.22 to + 22) have values that are (approximately) not significantly different from zero—or chance targeting of services.

example, if the students are ranked by need and the n most needy are not all in the sample of size n that is selected, then the method considers all possible selections of n students that would result in a set of students less needy than those actually selected.

Although in concept the method deals with ranks of students, in fact, the procedure compares mean neediness scores of the two subsets, and a variance term is derived from the concept of all possible subsets. Then a Z statistic is computed and compared to a Z table to obtain a value for *alpha*. For this procedure to have a theoretical basis, it is necessary to assume that the difference between the means, considered as a variable over all possible partitions of the N students into subsets of sizes n and N-n, is approximately normally distributed.

Beaton's suggested index considers neediness as a variable of at least ordinal character, rather than a dichotomous classification, and the ordering is specific to each school. The value of the index would be reduced if, for example, we decided to remove the most needy student from the participating group and substitute the student who ranks tenth in neediness, even though both might be sufficiently needy to qualify for CE. Thus, the index deals with finer distinctions than the dichotomous indexes discussed in this report, and more specifically addresses the within-school allocation of selection and services.

There are also some serious disadvantages inherent in this index, however. First, the difference between the means of the two subsets is not necessarily distributed normally; some tests of the method on artificial data sets revealed that the Z statistic it generates is sensitive to a monotonic transformation of the data, and to the shape of the distribution. These characteristics seem undesirable in light of the rationale for the statistic, which is based entirely on a ranking of students—ranks being, of course, unaffected by a monotonic transformation. A second disadvantage is that the index does not have a convenient and bounded range, as the previous indexes had. This makes interpretation less simple, since the theoretical range extends infinitely in both the positive and negative directions. It is not obvious whether a value of 2.0, for example, is large or small. While it is possible to express the unlikelihood of obtaining a given value of Z by chance, the statistic is not an easy-to-read indicator of degree of association.

Table 7-21 contains the distribution of Beaton's index values, by school and by grade within school, for reading and math. For reading, the school-level index values vary between +1.8 and −0.4, with only one percent of the values being less than or equal to zero (Since the mean of the CE group is subtracted from the mean of the non-CE group, the 'good' values of the index are the positive ones.) The distribution is roughly bell-shaped, and is centered above zero (with a median of +1.1). The school-level distribution for math is very similar, but shows somewhat poorer selection, with 91.3 percent of the values positive and a median value of +0.9. The index agrees with the previous ones in that it shows somewhat worse targeting for math than for reading.

A modified version of Beaton's index is also presented here. In keeping with the ranking rationale from which the statistic was developed, and in response to its sensitivity to monotonic transformations of the data, the statistic was also computed on the ranks of the

Table 7-21

Cumulative Percentage Distributions by School and Grade: Beaton's Index

Index Value	School	Grade 1	Grade 2	Grade 3	Grade 4	Grade 5	Grade 6
Reading							
	(N = 205)	(N = 143)	(N = 169)	(N = 172)	(N = 164)	(N = 156)	(N = 144)
Less than −1.0	0.0	3.5	0.6	0.6	0.0	0.0	0.0
−1.0 - −0.8	0.0	4.9	1.2	1.2	0.0	0.0	0.0
−0.8 - −0.6	0.0	7.0	1.2	1.2	0.0	0.0	0.0
−0.6 - −0.4	0.5	9.8	1.8	1.7	0.0	0.6	0.0
−0.4 - −0.2	0.5	12.6	2.4	2.9	0.6	0.6	2.1
−0.2 - −0.0	1.0	24.5	3.6	4.1	1.8	1.3	2.1
0.0-0.2	3.9	35.7	7.1	5.2	6.1	2.6	4.9
0.2-0.4	11.7	54.5	9.5	7.0	10.4	3.8	8.3
0.4-0.6	19.0	70.6	17.2	15.1	17.7	10.3	11.8
0.6-0.8	29.8	83.9	27.2	21.5	28.0	17.9	20.1
0.8-1.0	50.7	94.4	37.3	31.4	42.1	32.7	36.8
1.0-1.2	79.5	97.2	50.9	50.6	58.5	44.2	59.0
1.2-1.4	92.7	97.9	72.8	68.6	80.5	65.4	75.7
1.4-1.6	99.5	98.6	88.8	84.9	93.3	80.1	86.1
1.6-1.8	100.0	99.3	94.1	93.0	100.0	90.4	96.5
1.8-2.0		100.0	98.2	95.9		96.2	97.9
Math							
	(N = 161)	(N = 86)	(N = 96)	(N = 109)	(N = 108)	(N = 108)	(N = 105)
Less than −1.0	0.0	1.2	2.1	3.7	0.0	0.0	0.0
−1.0 - −0.8	1.2	4.7	2.1	3.7	0.9	0.0	1.0
−0.8 - −0.6	1.9	9.3	2.1	5.5	1.9	0.0	1.9
−0.6 - −0.4	2.5	12.8	3.1	9.2	3.7	0.0	1.9
−0.4 - −0.2	5.6	17.4	3.1	10.1	3.7	0.0	2.9
−0.2 - −0.0	8.7	23.3	6.2	11.0	7.4	0.9	4.8
0.0-0.2	11.2	34.9	15.6	13.8	11.1	6.5	5.7
0.2-0.4	20.5	43.0	24.0	18.3	19.4	9.3	16.2
0.4-0.6	30.4	59.3	32.3	23.9	30.6	14.8	31.4
0.6-0.8	47.2	72.1	45.8	34.9	43.5	29.6	42.9
0.8-1.0	67.1	79.1	61.5	46.8	57.4	42.6	56.2
1.0-1.2	80.7	88.4	79.2	67.9	75.0	60.2	73.3
1.2-1.4	91.3	97.7	89.6	77.1	83.3	76.9	88.6
1.4-1.6	95.0	98.8	94.8	84.4	91.7	90.7	96.2
1.6-1.8	98.1	100.0	99.0	92.3	97.2	92.6	98.1
1.8-2.0	98.8		99.0	96.3	98.1	96.3	100.0

neediness scores (instead of their means), within the school or within the grade at a school, derived from the CTBS scores. The distributions are shown in Table 7-22 for reading and math. No appreciable differences in the index values are apparent. The median index value for reading is +1.1 and the median for math is +0.9, the same as that obtained for the unmodified version, discussed above.

Thus, for the data distributions with which we are dealing, the version based on percentile scores yields essentially the same results as the modified version from ranks. Being simpler to compute, it is, therefore, preferable to the method based on within-school ranks. Neither approach yielded median Z values that are significantly different from zero, but that likely reflects the state of targeting more than the quality of the index.

Having decided to examine indexes based on ranks of participants and non-participants, we find it natural to include the Mann-Whitney U in the set of indexes to be considered. This statistic is a measure of the difference between the ranks of scores in two groups, and is, in fact, converted to a Z statistic for the final comparison. Like the other indexes already discussed in this section, therefore, it is a better measure of difference than of association. We can easily determine how unlikely any value of U would be, given the null hypothesis (and the assumed model), but it is not so straightforward to express the degree of agreement between the dichotomous variable and the ranked variable in terms of U. The Mann-Whitney U technique is described in detail by Siegel (1956).

Table 7-23 presents the distribution of the Z statistic derived from the Mann-Whitney U. (The Z values were selected for the index because their theoretical distribution is known, thus giving us some way to evaluate differences, and because they are comparable to the previous indexes in this section, which are also expressable as Z statistics.) The Z values present a more optimistic view of targeting than those derived by the previous method, because they have greater magnitude. The meaning of the Z statistic is not quite the same under the two methods, however: for the Mann-Whitney U it is an indication of the likelihood that the ranks observed in the two groups could have occurred if there were no real differentiation between the two groups, while the statistic suggested by Beaton deals with the likelihood that worse rankings could have occurred. The median values from the Mann-Whitney U approach are highly statistically significant for both reading and math.

Correlations with CTBS Ranks. The ranks of the CTBS reading and math scores can also be used to construct indexes through correlation, as the indexes of the first and second sections were constructed. The correlation coefficient constitutes an index with some useful characteristics, namely a bounded range and a center of zero. The use of the ranks of the CTBS scores for a correlational index preserves the advantages of the statistic suggested by Beaton and the Mann-Whitney U, as well. The ranks preserve more than a dichotomous distinction, treating need as more complex than mere presence or absence; they focus on within-school differences among students, and they disregard the differences among scores that, as argued previously in this section, are irrelevant to the concept of targeting. A correlation between the CTBS ranks and CE status (which is dichotomous) is a measure, then, of the agreement between the CE status and the ranks, indicating the degree to which the lowest achievers tend to be CE participants (and tend not to be non-participants).

Table 7-22

Cumulative Percentage Distributions by School and Grade: Beaton's Index Modified by Ranks

Index Value	School (N = 205)	Grade 1 (N = 143)	Grade 2 (N = 169)	Grade 3 (N = 172)	Grade 4 (N = 164)	Grade 5 (N = 156)	Grade 6 (N = 144)
Reading							
Less than −1.0	0.0	2.8	1.2	0.6	0.0	0.0	0.0
−1.0 - −0.8	0.0	4.2	1.2	0.6	0.0	0.0	0.0
−0.8 - −0.6	0.0	7.0	1.2	0.6	0.0	0.0	0.0
−0.6 - −0.4	0.5	10.5	2.4	1.7	0.0	0.6	0.0
−0.4 - −0.2	0.5	13.3	3.0	3.5	0.6	0.6	1.4
−0.2 - −0.0	1.5	24.5	3.6	4.7	1.8	1.9	3.5
0.0-0.2	4.9	34.3	7.1	5.8	6.1	2.6	6.2
0.2-0.4	12.7	54.5	10.7	9.9	10.4	4.5	9.7
0.4-0.6	19.5	69.2	17.8	16.3	17.7	11.5	11.8
0.6-0.8	30.2	85.3	27.2	21.5	28.0	19.2	22.2
0.8-1.0	50.7	95.1	36.1	30.2	42.1	32.7	38.2
1.0-1.2	85.4	97.9	52.1	55.2	58.5	46.2	57.6
1.2-1.4	98.0	97.9	77.5	74.4	80.5	74.4	83.3
1.4-1.6	100.0	99.3	97.0	95.3	93.3	93.6	95.1
Math							
Index Value	School (N = 161)	Grade 1 (N = 86)	Grade 2 (N = 96)	Grade 3 (N = 109)	Grade 4 (N = 108)	Grade 5 (N = 108)	Grade 6 (N = 105)
Less than −1.0	0.0	3.5	2.1	2.8	0.9	0.0	0.0
−1.0 - −0.8	1.2	4.7	2.1	5.5	0.9	0.0	1.9
−0.8 - −0.6	1.2	8.1	3.1	6.4	2.8	0.0	1.9
−0.6 - −0.4	2.5	11.6	3.1	8.3	3.7	0.0	1.9
−0.4 - −0.2	6.8	16.3	3.1	9.2	4.6	0.0	3.8
−0.2 - −0.0	9.9	26.7	5.2	10.1	8.3	0.9	4.8
0.0-0.2	13.7	36.0	15.6	12.8	12.0	5.6	8.6
0.2-0.4	21.7	44.2	25.0	19.3	20.4	8.3	18.1
0.4-0.6	31.7	58.1	33.3	25.7	30.6	16.7	32.4
0.6-0.8	47.2	72.1	42.7	34.9	44.4	31.5	43.8
0.8-1.0	68.3	80.2	62.5	51.4	57.4	42.6	57.1
1.0-1.2	86.3	89.5	81.2	66.1	73.1	63.9	76.2
1.2-1.4	94.4	96.5	91.7	84.4	90.7	82.4	90.5
1.4-1.6	98.1	100.0	97.9	95.4	98.1	93.5	96.2

Table 7-23

Cumulative Percentage Distributions by School and Grade: Z Statistics Based on Mann-Whitney U

Index Value	School (N = 205)	Grade 1 (N = 143)	Grade 2 (N = 169)	Grade 3 (N = 172)	Grade 4 (N = 164)	Grade 5 (N = 156)	Grade 6 (N = 144)
Reading							
0.0	0.5	8.4	0.6	0.6	1.8	1.3	2.8
0.5	2.0	28.0	5.3	5.2	6.1	2.6	5.6
1.0	3.9	51.7	13.0	11.6	12.2	7.1	13.9
1.5	6.3	72.0	24.3	21.5	27.4	17.3	23.6
2.0	8.8	84.6	33.1	33.1	40.9	30.8	32.6
2.5	13.7	90.9	44.4	43.6	53.7	42.3	52.8
3.0	18.0	93.7	56.8	57.0	65.9	58.3	64.6
3.5	24.4	96.5	66.9	68.0	79.9	69.9	74.3
4.0	30.7	99.3	76.3	82.0	85.4	80.1	86.1
4.5	37.6	99.3	82.8	86.6	93.3	86.5	87.5
5.0	43.4	100.0	91.1	92.4	96.3	90.4	90.3
5.5	52.2		94.1	96.5	97.6	93.6	92.4
6.0	61.5		97.6	98.3	98.8	96.8	95.1
6.5	65.9		98.8	98.8	98.8	97.4	97.9
7.0	73.7		98.8	99.4	100.0	98.7	97.9
7.5	79.5		100.0	100.0		100.0	98.6

Index Value	School (N = 161)	Grade 1 (N = 86)	Grade 2 (N = 96)	Grade 3 (N = 109)	Grade 4 (N = 108)	Grade 5 (N = 108)	Grade 6 (N = 105)
Math							
0.0	1.9	14.0	6.2	1.8	2.8	0.9	1.9
0.5	6.2	32.6	17.7	5.5	12.0	11.1	12.4
1.0	11.8	41.9	28.1	21.1	20.4	20.4	20.0
1.5	17.4	59.3	44.8	36.7	45.4	32.4	40.0
2.0	23.6	69.8	59.4	47.7	58.3	45.4	57.1
2.5	36.6	84.9	75.0	65.1	66.7	60.2	77.1
3.0	42.2	90.7	82.3	78.9	81.5	74.1	84.8
3.5	50.9	94.2	88.5	83.5	92.6	81.5	88.6
4.0	58.4	96.5	90.6	94.5	96.3	86.1	91.4
4.5	66.5	98.8	94.8	94.5	96.3	92.6	94.3
5.0	73.3	98.8	97.9	96.3	99.1	97.2	95.2
5.5	78.3	100.0	97.9	97.2	100.0	99.1	98.1
6.0	84.5		100.0	100.0		99.1	98.1
6.5	87.6					100.0	99.0
7.0	91.3						99.0
7.5	93.8						100.0

Note: Index values are midpoints of intervals of .5.

The cumulative percentage distributions for these correlations appear in Table 7-24 for reading and math. The reflected correlations are nearly all positive (at the school level, 98 percent for reading and 90.1 percent for math), but do not tend to be strongly so. The median value is about +0.37 for reading and about +0.24 for math. As with the previous indexes, then, there is evidence of better-than-chance targeting, but the agreement between CE participation and within-school achievement rank is not impressive.

The ranked CTBS scores were also correlated with the measures of amount of services received. While these service measures do not reflect CE participation, or even, necessarily, services that are specifically part of CE, the correlations of them with CTBS ranks indicate the degree of agreement between amount of instructional service received and the need for CE services. Table 7-25 contains the cumulative percentage distributions for the reflected correlations between the measure of need and the total exposure to instruction in reading and math. As with the previous indexes relating achievement to exposure, the reflected correlations tend to be positive, but only mildly so. These analyses confirm the earlier findings that exposure to services is not a variable that contributes to favorable targeting indexes.

The corresponding distributions for the index based on total resource cost for reading and math instruction appear in Table 7-26. The correlations for this variable show stronger relationships than those for the hours of services received. Most of the reflected correlations are positive (92.2 percent for reading and 82 percent for math), but the median values are not particularly strong (about +0.22 for reading and +0.16 for math). The indexes based on resource cost, then, are about as sensitive as those based on CE participation.

DISCUSSION OF ALTERNATIVE INDEXES

Throughout this section we have taken the position that no one index of targeting will serve for all situations against which targeting is to be assessed. What constitutes proper assignment of CE to students may be specified in various ways, depending on several considerations. Selection of a targeting index for any given evaluation problem, then, requires that the evaluator exercise judgment. For this reason, we have, with each of the 25 indexes, presented the rationale for its derivation and some empirical data on its distribution in the sample of SES schools. The purpose of this chapter is to summarize the discussion of the alternative indexes without the distractions caused by the concurrent development of the indexes in the previous sections.

In order to discuss the wide range of indexes, all 25 are named in the paragraphs below. Each index is also provided with an acronym that can serve as a mnemonic. The acronyms will be employed in this chapter for the sake of brevity in the tables that follow.

The first indexes developed,

1. P.BELOW.50 — CE participants scoring below the 50th percentile
2. P.BELOW.40 — CE participants scoring below the 40th percentile

Table 7-24

Cumulative Percentage Distributions by School and Grade: Reflected Point-Biserial Correlations Between CE Participation and Achievement Rank

Reading

Index Value	School (N = 205)	Grade 1 (N = 143)	Grade 2 (N = 169)	Grade 3 (N = 172)	Grade 4 (N = 164)	Grade 5 (N = 156)	Grade 6 (N = 144)
−.20 or less	0.0	7.0	1.2	1.2	0.0	0.6	0.0
−.15	0.0	9.1	2.4	1.7	0.0	0.6	0.7
−.10	0.5	11.2	2.4	2.3	0.6	0.6	0.7
−.05	1.0	18.2	3.6	4.1	1.2	1.3	1.4
.00	2.0	24.5	3.6	4.7	2.4	1.9	3.5
.05	4.4	32.9	5.3	5.2	4.9	1.9	6.2
.10	6.8	42.0	9.5	7.0	8.5	3.2	7.6
.15	12.2	59.4	13.6	9.9	12.8	7.7	13.9
.20	22.0	68.5	18.3	15.7	19.5	10.3	23.6
.25	28.3	76.2	24.9	20.9	25.0	17.3	29.2
.30	43.4	83.2	29.6	26.7	34.1	24.4	42.4
.35	54.6	90.9	34.9	33.7	43.9	32.7	54.9
.40	70.7	95.1	40.8	41.3	52.4	44.9	63.9
.45	84.4	97.2	49.1	54.1	65.2	59.6	70.8
.50	90.7	98.6	58.0	65.1	72.6	67.3	73.6
.55	96.1	98.6	71.6	76.2	79.9	78.8	84.0
.60	98.0	98.6	79.9	87.2	88.4	87.2	88.2
.65	99.0	98.6	85.8	90.7	95.1	90.4	94.4
.70	99.5	99.3	92.3	94.8	96.3	96.2	97.2
.75	100.0	99.3	98.2	97.7	97.6	96.8	97.9
.80		99.3	99.4	98.3	98.8	97.4	98.6

Math

Index Value	School (N = 161)	Grade 1 (N = 86)	Grade 2 (N = 96)	Grade 3 (N = 109)	Grade 4 (N = 108)	Grade 5 (N = 108)	Grade 6 (N = 105)
−.20 or less	0.6	8.9	2.1	5.5	2.8	0.0	1.0
−.15	1.9	10.5	3.1	6.4	3.7	0.0	1.0
−.10	3.7	14.0	3.1	9.2	3.7	0.0	1.9
−.05	6.8	17.4	3.1	9.2	6.5	0.0	3.8
.00	9.9	29.1	7.3	11.0	8.3	0.9	4.8
.05	15.5	39.5	16.7	13.8	11.1	1.9	12.4
.10	24.8	43.0	25.0	16.4	13.9	12.0	16.2
.15	36.6	48.8	35.4	22.9	29.6	21.3	29.5
.20	45.3	61.6	43.7	29.4	37.0	25.0	40.0
.25	62.7	72.1	51.0	44.0	48.1	37.0	56.2
.30	78.9	81.4	54.2	55.0	56.5	47.2	66.7
.35	88.2	89.5	64.6	64.2	69.4	57.4	75.2
.40	91.9	89.5	76.0	71.6	83.3	69.4	80.0
.45	95.7	91.9	84.4	78.9	88.0	75.9	84.8
.50	97.5	97.7	92.7	83.5	91.7	82.4	91.4
.55	98.8	97.7	96.9	89.0	95.4	88.9	94.3
.60	98.8	98.8	96.9	92.7	97.2	95.4	96.2
.65	98.8	98.8	97.9	94.5	100.0	99.1	98.1
.70	100.0	98.8	97.9	95.4		99.1	99.0
.75		100.0	99.0	97.2		99.1	99.0
.80			99.0	97.2		100.0	99.0

Note: Index values are midpoints of intervals of .05.

Table 7-25

Cumulative Percentage Distributions by School and Grade: Reflected Product-Moment Correlations Between Total Exposure to Instruction and Achievement Rank

Reading

Index Value	School (N = 206)	Grade 1 (N = 183)	Grade 2 (N = 183)	Grade 3 (N = 184)	Grade 4 (N = 180)	Grade 5 (N = 178)	Grade 6 (N = 162)
-.40 or less	0.0	4.4	6.0	2.7	3.3	2.8	6.2
-.35	0.0	7.7	8.7	6.0	4.4	5.1	7.4
-.30	1.0	12.0	12.0	8.7	6.7	6.7	8.6
-.25	2.9	17.5	18.6	14.1	8.9	7.9	11.7
-.20	6.3	23.0	20.8	15.8	13.3	11.8	17.9
-.15	9.7	29.0	27.3	20.1	14.4	15.7	22.8
-.10	14.6	38.3	33.9	25.5	23.3	23.0	32.7
-.05	26.7	45.9	37.7	28.8	30.0	30.9	36.4
.00	43.2	55.7	43.7	34.8	36.1	34.3	46.3
.05	58.3	65.6	49.7	39.7	40.6	39.1	51.2
.10	72.8	73.2	59.6	49.5	48.9	46.1	56.2
.15	83.5	80.9	64.5	56.0	55.0	52.2	64.8
.20	87.9	88.5	69.9	63.6	59.4	62.9	67.9
.25	93.2	90.2	72.1	69.6	63.3	68.5	71.6
.30	95.6	92.3	76.0	75.0	68.9	73.0	79.0
.35	97.1	95.6	81.4	81.0	73.9	80.3	81.5
.40	98.5	97.3	85.8	83.2	77.2	83.1	84.0
.45	99.5	97.3	89.1	88.0	81.7	87.6	86.4
.50	99.5	97.3	89.6	91.3	87.8	89.3	92.6
.55	99.5	97.8	90.7	93.5	90.0	91.6	95.1
.60	99.5	98.4	93.4	95.7	93.3	93.8	96.3

Math

Index Value	School (N = 206)	Grade 1 N = 183)	Grade 2 (N = 183)	Grade 3 (N = 183)	Grade 4 (N = 180)	Grade 5 (N = 177)	Grade 6 (N = 162)
-.40 or less	1.0	5.5	9.8	7.7	6.7	5.1	3.7
-.35	1.0	5.5	13.7	8.7	10.6	7.9	6.8
-.30	2.4	10.4	16.4	11.5	11.7	12.4	8.6
-.25	5.8	13.7	20.2	18.0	16.1	16.9	17.3
-.20	9.7	21.3	23.5	24.0	21.7	26.0	24.1
-.15	18.0	31.1	32.2	30.6	30.6	33.3	37.0
-.10	31.6	40.4	39.9	39.3	40.6	41.8	48.8
-.05	45.1	53.6	47.0	47.5	50.0	50.3	59.3
.00	65.0	62.3	57.4	54.6	59.4	56.5	67.3
.05	77.2	67.2	63.9	63.4	66.1	63.8	75.9
.10	89.8	76.0	75.4	68.9	71.1	68.9	80.2
.15	92.2	80.3	82.0	79.2	77.2	73.4	82.7
.20	94.7	87.4	86.3	83.1	82.8	77.4	88.9
.25	97.6	90.7	91.8	86.9	88.9	83.1	89.5
.30	99.0	93.4	92.9	90.2	89.4	86.4	92.6
.35	99.0	96.2	96.7	93.4	92.2	89.3	96.3
.40	99.5	97.8	97.3	94.0	96.7	92.1	97.5
.45	99.5	99.5	97.3	94.0	98.3	93.8	97.5
.50	100.0	99.5	97.3	95.6	98.3	94.9	97.5
.55		100.0	97.3	97.3	98.3	94.9	97.5
.60			98.4	98.4	98.3	96.4	98.1

Note: Index values are midpoints of intervals of .05

Table 7-26

Cumulative Percentage Distributions by School and Grade: Reflected Product-Moment Correlations Between the Resource Costs of Instruction and Achievement Rank

Reading

Index Value	School (N = 206)	Grade 1 (N = 183)	Grade 2 (N = 183)	Grade 3 (N = 182)	Grade 4 (N = 180)	Grade 5 (N = 178)	Grade 6 (N = 161)
-.30 or less	1.0	7.7	3.8	7.1	2.2	3.4	4 3
-.25	1.0	7.7	4.9	7.7	3 9	3.4	5 6
-.20	1.0	11 5	4.9	7.7	6.1	5.1	6.8
-.15	1.5	19 7	6.6	8.8	7 8	8.4	8.7
-.10	3.9	25.1	8.7	10.4	8.9	9.6	9.9
-.05	7.8	34.4	11.5	11.5	12.2	11 2	13.7
.00	11.7	42 1	14.8	14.8	13 9	11.8	17.4
.05	17.0	50 8	19 1	15.9	17 2	14.6	22.4
.10	23 3	59 0	21 3	18 1	21 7	18.0	26.7
.15	38.8	66 1	23 5	25 3	24.4	21.3	29.8
.20	54 9	74.3	29 5	30.8	26.1	25.8	36.6
.25	68 4	79.2	35 0	36.3	32.8	33.7	46.6
.30	80.1	86 9	41 0	40.1	38.9	41.6	51 6
.35	86 9	91 3	45 4	47 3	49.4	50.0	61.5
.40	93 2	94.5	54.6	50 0	55.0	59.0	67.7
.45	96 6	96 7	62.8	59.3	61 7	67 4	73 3
.50	98 5	97 8	68.9	70 3	76.1	73 6	78.9
.55	99.0	98 9	77 0	78.6	82.8	79.8	87.0
.60	99 0	99.5	86.3	88.5	88.3	86 5	91.3
.65	99 5	100 0	91 8	92.9	93.9	91.6	96.3
.70	100 0		96.2	95.6	97 2	93.8	96.9

Math

Index Value	School (N = 206)	Grade 1 (N = 183)	Grade 2 (N = 183)	Grade 3 (N = 181)	Grade 4 (N = 180)	Grade 5 (N = 177)	Grade 6 (N = 161)
-.30 or less	1.0	7.7	12.0	11.6	8.9	7.3	6.8
-.25	1.5	9.8	15 8	14 9	10.6	12.4	8.9
-.20	3 9	14 8	18.6	18.2	13.3	16.9	11.2
-.15	6.8	17.5	21 9	21 5	17.2	18.1	13.7
-.10	11 2	21 9	24.6	24.9	20.0	24.3	16.8
-.05	18 0	31 1	27 9	27.6	23.9	28.2	22.4
.00	29.6	41 5	37.2	33 1	28.9	32.2	27.3
.05	42 7	47.0	43.2	39 2	38 9	40.7	34.8
.10	52 4	57.9	48 6	45.9	45.0	44.1	38.5
.15	65 5	63 4	60 1	53.0	51.7	50.3	45.3
.20	74.3	73 2	65 6	59.7	62 8	54.8	53.4
.25	87.4	76.0	71.0	67.4	72.2	63.8	64 6
.30	92.7	83.6	76 0	72 9	77.8	67.2	73.9
.35	96 6	89 1	81.4	79.6	82.2	74.0	83.2
.40	98.1	91.8	87.4	84.0	85.6	78.5	87.0
.45	99.5	94 5	91 3	87 8	89.4	84.2	91.3
.50	100 0	97.3	94.0	90.6	92.8	88.1	93.8
.55		98.4	95.6	92.3	96.7	93.2	96.9
.60		99.5	96.7	95.6	99.4	96.6	98.1
.65		99.5	97.3	96.1	100.0	98.9	98.8
.70		99.5	98.4	97.2		99.4	99.4

Note Index values are midpoints of intervals of .05

150

3. P.BELOW.35 — CE participants scoring below the 35th percentile

could, if computed on a cutoff acceptable to the evaluator, minimally indicate the degree to which schools comply with the simplest interpretation of Title I regulations. They are not recommended for any other purpose, however, because of their insensitivity to the neediness of students who are not participating in CE.

The unadjusted *phi* coefficients of the second section,

4. PHI.CES.50 — *Phi* coefficient between CE participation and achievement scores dichotomized at the 50th percentile

5. PHI.CES.40 — *Phi* coefficient between CE participation and achievement scores dichotomized at the 40th percentile

6. PHI.CES.35 — *Phi* coefficient between CE participation and achievement scores dichotomized at the 35th percentile

7. PHI.CES.TJ — *Phi* coefficient between CE participation and teacher's judgment of need for CE

are most appropriate for the assessment of targeting when the entire process of allocating CE to students, rather than just within-school CE participation, is of interest; when need for CE is seen as a dichotomous, nationally-normed variable, and, when it is participation rather than amount of CE service that is measured.

When the *phi* coefficient is adjusted for unequal marginals,

8. CPHI.CE.40 — Corrected *phi* coefficient between CE participation and achievement scores dichotomized at the 40th percentile

9. CPHI.CE.35 — Corrected *phi* coefficient between CE participation and achievement scores dichotomized at the 35th percentile

10. CPHI.CE.TJ — Corrected *phi* coefficient between CE participation and teacher's judgment of need for CE

the focus is more sharply on within-school-targeting, but the other considerations still apply.

When dichotomous indicators of need are correlated with level of services,

11. R.INSTR.40 — Correlation between total hours of instruction received and achievement scores dichotomized at the 40th percentile

12. R.INSTR.35 — Correlation between total hours of instruction received and achievement scores dichotomized at the 35th percentile

13. R.INSTR.TJ — Correlation between total hours of instruction received and teacher judgment of need for CE

14. R.RCOST.40 — Correlation between resource-costs of instruction and achievement scores dichotomized at the 40th percentile

15. R.RCOST.35 — Correlation between resource-costs of instruction and achievement scores dichotomized at the 35th percentile

16. R.RCOST.TJ — Correlation between resource-costs of instruction and teacher's judgment of need for CE

the resulting indexes address somewhat different questions. Rather than the targeting of CE to needy students, these indexes measure the targeting of *instructional services* to the needy students. (It is important to note that these analyses do not identify CE instructional services specifically; indeed, it is not possible to differentiate accurately which services are associated with what programs.) These indexes would be most appropriate when the question is whether needy students are receiving the most instructional services, disregarding any labels associated with being selected for a CE program. The major disadvantage with these indexes arises from the difficulty in measuring the instructional services and the resource-costs.

The indexes from the third section are all correlations between achievement percentile scores and either CE participation rates or levels of instructional service.

17. R.CE.PERCN — Correlation between CE participation and achievement percentile scores

18. R.CE.INSTR — Correlation between CE participation and total hours of instruction received

19. R.CE.RCOST — Correlation between CE participation and resource-costs of instruction

Like the indexes based on dichotomizations in the second section, these examine association between neediness as defined by national norms, and CE participation rates or levels of service. They differ only insofar as the percentile scores represent a finer ordering, nationally, than the dichotomous versions, and so they are appropriate when the dichotomy-based indexes are not useful or desired. In practice, these percentile-based indexes would be easier to compute because the percentiles have to be obtained in either case, and using them directly saves the effort of dichotomizing before computing the correlation coefficients. The major conceptual disadvantage of these indexes is that they indicate only the strength of the linear relationship between CE participation rates or service levels and achievement percentiles. This seems to imply that a five-point difference between two scores has the same practical significance (in terms of CE participation) at all points along the score range.

Finally, the indexes from section 4,

20.	IND.BEATON	Index suggested by Albert Beaton
21.	MOD.BEATON	Modification of the index suggested by Beaton
22.	U.CE.ARANK	Mann-Whitney U between CE participation and achievement-score rank
23.	R.CE.ARANK	Correlation between CE participation and achievement-score rank
24.	R.INS.RANK	Correlation between total hours of instruction received and achievement-score rank
25.	R.CST.RANK	Correlation between resource-costs of instruction and achievement-score rank

are based on within-school ranking of students by their achievement levels. These indexes are particularly expensive to compute for a large sample because they entail ranking students within each school. For a school evaluation, however, the use of local rank ordering focuses on targeting within the school in the most rigorous way, and the added expense to compute ranks for a single school is probably not of great consequence.

Sensitivity of Indexes. In addition to knowing which indexes are conceptually appropriate for a given application, the prospective user would benefit from some indication of the sensitivity of the index. As each index was presented in the previous sections, a cumulative frequency distribution was provided. The distributions show the central tendencies of the indexes, their scatter, and the general shape of their distributions. It is tempting to consider the scatter of the distribution as an indication of sensitivity, since greater variation does indicate that the schools are being differentiated more. Two caveats should be kept in mind, however: first, it would be misleading to compare the ranges of indexes with known sampling distributions that differ, such as the Z statistic and most correlation coefficients; second, the differentiation of schools may be in large part caused by factors irrelevant to the concerns about within-school targeting (i.e., they may reflect to some degree the level of CE funding in a school, or the way the school focuses its CE services at different grades).

It is also appealing to think of the central tendency as an indicator of sensitivity. If we know *a priori* that targeting is excellent in one set of schools and is very poor in another set, then we would prefer the index that best differentiates between the sets. Similarly, if we knew that all schools have excellent targeting, we would prefer the index for which targeting seems best. Even if we only knew that schools tend to have good targeting, the same decision rule would apply. Unfortunately, we have no such knowledge, and an index that shows schools in the best light is not necessarily better than one that presents a less optimistic picture.

On the basis of our work with the indexes, however, we feel that most schools are attempting to honor the intentions of CE programs in allocating CE status, services, or costs to those students who are defined as needing them. Misallocations do occur, either by error in determining need or by interpretations of guidelines and regulations of the CE program that militate against a perfect relationship between the need and the status, service, or cost. Such errors or misinterpretations, although numerous, do not appear to us to outnumber the instances of good targeting.

Empirical Test of Validity of the Indexes. In order to obtain some basis for inferring the validity of the indexes, we followed an approach similar to the methods used in validating some tests and test items. We first assumed that there was a sensible basis for all, or most, of the indexes. Therefore, we could treat an index as valid to the degree that it correlated strongly with the other indexes. The matrix of intercorrelations among all 25 indexes is presented in Table 7-27, the coefficients based on indexes for reading above the diagonal and those based on indexes for math below the diagonal.

The correlation coefficients in Table 7-27 have surprisingly low values. This indicates that our indexes are not reflecting the same aspects of targeting. From inspection of the matrix of coefficients we can see that the indexes for reading are interrelated quite similarly to the indexes for math. Further, the clusters of high coefficients are also similar for the two subject areas. Four clusters appear to be consistent and strong, likely accounting for most of the common variance in a factor-analytic sense. The clusters are:

A Coefficients based on *phi*, on Beaton's procedures, on the Mann-Whitney U, and on correlations with achievement ranks. This is the largest cluster of coefficients, and all the variables have the least incidence of negative correlations with the other variables in the matrix. It follows, then, that other considerations aside, a favored index should be selected from this cluster. The cluster includes indexes numbered 4, 5, 6, 7, 17, 20, 21, 22, and 23.

B Coefficients based on only the participants below certain cutoffs and those based on corrected *phi* coefficients. This cluster appears to have some common base in the mere fact that they share the same numerical component. The cluster includes indexes numbered 1, 2, 3, 8, and 9.

C Coefficients based on total hours of instruction are highly intercorrelated. The common aspect, of course, is that hours of service is the component. It also follows, then, that indexes based on hours of service are not particularly similar to (or interchangeable with) those based on CE participation. The cluster includes indexes numbered 11, 12, 13, 18, and 24.

D Although moderately related to the cluster based on hours of service, those based on resource-costs (a transformation of hours of service) tend to cohere into a cluster by themselves. We observe the phenomenon, then, that these indexes are not interchangeable with those based on services or on CE participation. The cluster includes indexes numbered 14, 15, 16, 19, and 25.

Table 7-27

Correlations Among 25 Targeting Indexes for Schools, Based on Over 200 Schools
(Coefficients above diagonal for reading; below diagonal for math. Decimal points omitted.)

Indexes	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
1. S.BELOW.50		90	84	16	24	24	12	66	53	25	−03	−05	04	05	03	06	16	−01	03	13	19	14	19	02	08
2. S.BELOW 40	92		95	−03	17	19	−03	69	56	15	−07	−07	−02	01	−01	−01	04	−04	−01	01	06	04	07	−01	04
3. S.BELOW.35	89	97		−09	12	22	−07	63	58	10	−09	−08	−05	−01	00	−02	01	−07	−02	−01	03	04	04	−05	03
4. PHI.CES.50	37	24	21		91	86	73	35	38	34	26	20	16	39	38	30	95	23	37	74	76	63	94	22	36
5. PHI.CES.40	38	35	31	94		94	70	53	52	37	28	23	17	41	39	30	94	26	38	74	77	63	93	25	37
6. PHI CES.35	39	37	38	91	97		66	51	59	32	27	23	15	42	43	30	93	24	39	74	76	61	91	23	38
7. PHI.CES TJ	22	12	09	72	68	68		27	27	72	29	23	26	41	40	43	74	25	38	60	63	58	74	24	37
8. CPHI CE.40	70	76	76	40	53	55	26		91	47	12	07	05	25	21	11	42	12	22	58	62	31	43	14	23
9. CPHI.CE.35	66	71	76	36	48	55	21	95		41	14	09	09	24	22	10	48	15	21	61	65	30	49	15	22
10. CPHI CE TJ	43	39	36	36	32	32	59	30	23		21	16	20	29	28	29	36	18	26	49	52	32	37	18	26
11. R.INSTR.40	−03	−01	03	04	08	11	15	02	04	−03		95	55	52	51	23	27	95	52	28	28	19	26	93	50
12. R.INSTR.35	−01	−01	04	04	06	11	15	03	05	−03	95		56	50	53	24	21	93	51	23	23	17	20	92	50
13. R.INSTR.TJ	02	−03	−04	31	28	27	34	−03	−05	18	48	48		15	17	51	19	58	18	11	11	10	19	59	17
14. R.RCOST.40	09	14	14	21	28	31	29	25	24	15	48	49	17		96	59	37	47	96	45	46	30	36	45	95
15. R.RCOST.35	09	13	15	20	27	31	28	26	26	12	48	52	15	97		59	35	48	94	40	41	27	34	46	93
16. R.RCOST TJ	15	11	09	42	43	44	51	19	18	29	37	37	51	62	59		31	23	59	26	27	25	30	22	60
17. R CE.PERCN	35	27	25	96	96	95	72	45	43	36	09	08	30	27	26	46		24	36	77	80	67	99	23	34
18. R.CE.INSTR	−02	−02	01	05	06	10	16	−02	00	−02	95	92	52	43	43	37	09		53	25	25	14	24	99	52
19. R.CE.RCOST	12	15	15	22	28	31	32	25	24	16	48	49	21	95	94	64	29	48		42	42	26	35	51	99
20. IND.BEATON	55	52	53	59	61	64	43	77	75	47	−05	05	13	28	28	30	63	03	29		99	58	73	22	38
21. MOD.BEATON	58	54	55	64	66	68	48	79	78	50	05	04	13	30	29	34	69	02	31	98		60	78	23	39
22. U CE.ARANK	16	12	10	63	66	64	58	27	24	27	07	05	13	25	23	32	71	07	26	41	47		67	13	25
23. R.CE.ARANK	36	28	26	95	95	94	72	46	43	36	08	08	29	27	26	46	99	08	29	60	68	71		23	34
24. R.INS RANK	−01	−02	02	04	06	08	16	−01	00	−01	94	91	51	41	42	36	08	99	47	03	03	07	08		51
25. R.CST.RANK	13	16	16	22	27	29	34	25	23	17	47	47	19	94	92	62	28	47	99	27	30	27	29	47	

Table 7-28

Attributes of Each of 25 Targeting Indexes

Index	Is the index easy to calculate?	Does the index consider the receipt of services	Does the index consider 'selection' for CE only?	Is the index based on national or school 'norms'?	Does the index consider all the 'needy' students?	Are schools penalized if they 'spread' CE to all students?	Are schools penalized if they target CE to selected grades only?	Does the index have a known sampling distribution?
1. S.BELOW.50	Yes	No	Yes	National	No	Yes	No	No
2 S.BELOW 40	Yes	No	Yes	National	No	Yes	No	No
3. S.BELOW.35	Yes	No	Yes	National	No	Yes	No	No
4. PHI.CES.50	Yes	No	Yes	National	Yes	Yes	Yes	Yes
5 PHI CES.40	Yes	No	Yes	National	Yes	Yes	Yes	Yes
6. PHI.CES 35	Yes	No	Yes	National	Yes	Yes	Yes	Yes
7. PHI CES.TJ	Yes	No	Yes	School	Yes	Yes	Yes	Yes
8 CPHI CE.40	Yes	No	Yes	National	Yes	Yes	Yes	Yes
9. CPHI.CE 35	Yes	No	Yes	National	Yes	Yes	Yes	Yes
10. CPHI.CE.TJ	Yes	No	Yes	School	Yes	Yes	Yes	Yes
11. R.INSTR.40	No	Yes	No	National	Yes	Yes	Yes	Yes
12 R.INSTR.35	No	Yes	No	National	Yes	Yes	Yes	Yes
13 R.INSTR.TJ	No	Yes	No	School	Yes	Yes	Yes	Yes
14. R.RCOST.40	No	Yes	No	National	Yes	Yes	Yes	Yes
15. R.RCOST 35	No	Yes	No	National	Yes	Yes	Yes	Yes
16. R.RCOST TJ	No	Yes	No	School	Yes	Yes	Yes	Yes
17. R CE.PERCN	Yes	No	Yes	National	Yes	Yes	Yes	Yes
18. R CE.INSTR	No	Yes	No	—	—	Yes	Yes	Yes
19 R CE.RCOST	No	Yes	No	—	—	Yes	Yes	Yes
20 IND.BEATON	No	No	Yes	School	Yes	Yes	Yes	Yes
21. MOD BEATON	No	No	Yes	School	Yes	Yes	Yes	Yes
22. U.CE.ARANK	Yes	No	Yes	School	Yes	Yes	Yes	Yes
23. R.CE ARANK	Yes	No	Yes	School	Yes	Yes	Yes	Yes
24 R.INS.RANK	No	Yes	No	School	Yes	Yes	Yes	Yes
R.CST.RANK	No	Yes	No	School	Yes	Yes	Yes	Yes

E. We expected a cluster of indexes based on teachers' judgments of need for CE, but such a cluster does not emerge with any strength. It seems that whatever unique information teachers' judgments bring to targeting is submerged in other components of the indexes.

Recommendations. Finally, we wish to provide the reader with some recommendations regarding which index might be preferred. As the body of this report indicates, no recommendations can be made without consideration of the circumstances in which the index is to be calculated or employed. Therefore, by way of summary, we present Table 7-28, which lists each of the indexes and indicates its attributes. Based on the entries of this table, the reader or user should be able to arrive at an informed judgment about the relative merits of each index.

CHAPTER 8. TEACHER JUDGMENT OF NEED FOR COMPENSATORY EDUCATION

Charles E. Kenoyer

Teacher judgments of student need for compensatory education (CE) agree moderately with scores on the CTBS. If a CTBS 'cutting score' at the 35th percentile is adopted, a near-maximum agreement rate (for about three fourths of the students) is obtained. The accuracy of teacher judgments is not influenced by the racial/ethnic or economic characteristics of the students, the extent of individualization of instruction, how students are selected for compensatory services, or teacher training and experience.

Teacher judgment is one of the permissible ways for schools to select students for CE, and is widely used (Wang, Hoepfner, Zagorski, Hemenway, Brown, and Bear, 1978). It is important, therefore, to know how well these judgments agree with objective measures of student performances. The Comprehensive Tests of Basic Skills (CTBS) reading and math subtests constitute the best measures of achievement that are available in the Sustaining Effects Study (SES), and so were used in these analyses as the criteria against which the judgments were gauged. The first step was to develop an index of agreement between teacher judgment of student need for CE and the CTBS reading and math scores.

SAMPLE SELECTION

Teachers and students were selected for these analyses as follows. 200 teachers were randomly selected from the SES teacher file at each grade. (Teachers associated with more than one grade were deleted.) A file linking teachers to their homeroom students was then used to select all students for each teacher, because it was the homeroom teacher who provided the judgments regarding each student's need for CE.

A teacher was subsequently included in any analysis except where one or more of the variables were missing. Many of the teachers taught no reading or no math, and so were excluded from any analysis requiring such data. Those who taught both subjects were included in the sample for each kind of analysis.

THE TEACHER JUDGMENTS

The Student Background Checklist item, in which the teacher judges each student's need for both reading and math CE, was coded to yield to dichotomous classification decisions (whether the student needs reading CE or not, and similarly for math), which are treated independently (This approach is not based on an assumption that the two decisions are statistically independent of each other, it is known from previous analyses that the judgments of student need for the two subjects are highly correlated. The present analyses, however, call for separate consideration of the two school subjects to simplify interpretations of the association between teacher judgment and test results for each.) The analyses for reading and math are based on nearly identical samples of teachers and students. Results for the two subjects are never analyzed jointly, but in parallel.

The measure of agreement selected for the analyses is the correlation between a teacher's dichotomous classification of students by need for compensatory services in one of the subjects and the CTBS score for the same subject. Because it would be both inconvenient and costly to generate all the correlations that were needed—every correlation for each teacher in the sample—an approximate method was used. For each of the variables to be correlated, the sum and sum of squares were computed. If a missing value was encountered, the observation was deleted. For each pair of variables to be correlated, the sum of crossproducts was computed, deleting any observation for which one of the pair was missing. This procedure leads to different subsamples for the terms of the correlation formula, hence its departure from an exact correlation. It was assumed that each of the subsets was unbiased and representative of the whole sample, i.e., that the missing observations were not systematic. For each pair of variables x and y, then, the correlation was computed as:

$$\frac{\frac{\Sigma xy}{N_{xy}} - \left(\frac{\Sigma x}{N_x}\right)\left(\frac{\Sigma y}{N_y}\right)}{\sqrt{\frac{\Sigma x^2}{N_x} - \left(\frac{\Sigma x}{N_x}\right)^2}\sqrt{\frac{\Sigma y^2}{N_y} - \left(\frac{\Sigma y}{N_y}\right)^2}}$$

The resulting correlations were transformed to Fisher's Z for the subsequent steps in the analysis.

We have no reason to expect that the validity of either teacher judgment of student need or our achievement measure is uniform over all grades, so the analyses were performed by grade throughout this report. Means of the correlations and their Z transforms (used for further calculation) are presented in Table 8-1 for both reading and math. The transformed coefficients are called 'Teacher Judgment Index for Reading' (TJIR) and 'Teacher Judgment Index for Math' (TJIM) and serve as indexes of teacher accuracy. Since teacher judgments indicate need, and CTBS scores indicate achievement level, agreement is indicated by negative correlations That is, students who are judged to be needy tend to have the lowest CTBS scores.

It is apparent that the agreement between teacher judgment and CTBS scores is poorest at the first grade, where teachers had little opportunity to become familiar with the students or their backgrounds and where achievement test scores can be expected to be least valid. This holds true for both reading and math. Although there is some variability over the other grades, it is small in comparison to the differences between first and second grades.

For many of the teachers in the sample, the correlations described above could not be computed because the teacher judged all students the same, and therefore there was no variance for that teacher on the teacher-judgment index. It was necessary to omit those teachers from the computation of the correlations reported in Table 8-1. The numbers and percentages of teachers omitted are shown by grade in Table 8-2.

Table 8-1

Teacher Judgment of Student Need and Achievement Scores: Mean Correlations and Transformed Indexes, by Grade

Grade	Reading		Math	
	Correlation	TJIR	Correlation	TJIM
1	-.23	-.26	-.33	-.37
2	-.60	-.75	-.43	-.50
3	-.55	-.68	-.49	-.57
4	-.53	-.65	-.42	-.49
5	-.55	-.68	-.49	-.59
6	-.54	-.65	-.44	-.51
Total Sample	-.50	-.61	-.44	-.51

Table 8-2

Number and Percentages of Teachers for Whom Correlations Could Not Be Computed, by Grade

Grade:	1	2	3	4	5	6	Whole Sample
			Reading				
Number	27	19	12	18	16	19	111
Percentage	16	11	7	10	9	15	11
			Math				
Number	37	36	18	26	25	16	158
Percentage	22	22	10	16	16	12	17

This finding suggests that student-level correlations, by teachers, may underestimate the degree of agreement. Underestimation would result if a considerable number of the teachers had only students who correctly fall into one teacher-judgment category, as would tend to occur, for example, in cases of ability grouping. Such clustering of high- or low-ability students would eliminate from consideration large numbers of students whose achievement scores are consistent with their teachers' judgments. These students are not eliminated when correlations are computed over the entire sample, rather than by teacher (but still subset by grade). For this analysis, teachers are disregarded, and only one correlation is computed for reading, and one for math, at each grade. The correlations appear in Table 8-3.

Table 8-3

Correlations Between Teacher Judgment and CTBS Scores at the Teacher Level, by Grade

Grade:	1	2	3	4	5	6
Reading	−.22	−.60	−.62	−.57	−.59	−.56
Math	−.31	−.42	−.51	−.45	−.49	−.43

Some of the correlations at the higher level of aggregation (that is, not at the teacher level) are slightly larger than those computed at the teacher level, but the majority are somewhat smaller. For reading, the pooled (not-by-teacher) correlations are consistently larger, with only one exception at the first grade, but again the differences are small. This result suggests that the within-teacher correlations are, on the average, adequate representations of the association between teacher judgment and CTBS scores, and so can serve as a valid index of agreement.

PERFORMANCE CURVES FOR THE TWO TEACHER JUDGMENT CATEGORIES

Given that the agreement between teacher judgments and CTBS scores is less than perfect, it is of interest to know whether there is some level of performance on the CTBS that corresponds to the boundary at which teachers make different judgments, and how well teacher judgments correspond to a division made at such a level. In order to examine these questions, students at each grade were divided into two groups according to whether they were judged by their teachers as needing CE in reading or math, or as not needing it. The cumulative percentage of students in each group was plotted against percentile scores for each grade and subject, as shown in Figures 8-1 and 8-2. (Every fifth percentile point is plotted.)

It is apparent that judgment groups for both reading and math have less separation for first grade than for the other grades, confirming the relative sizes of correlations for the grades described previously. If first grade is disregarded, it is apparent that the two judgment-based groups are widely separated in the plot at about the 35th achievement percentile. For example, the classification of second-grade math students on the basis of scoring above or below the 35th percentile would be consistent with the teacher-judgment classification for 80 percent of those judged not needy and about 70 percent of those judged needy. For many of the other subject-grade combinations, the agreement would be higher. In fact, all the reading curves are more widely separated than the math curves.

This finding indicates that using a point near the 35th percentile as a cutting point would optimize agreement between teacher judgment and CTBS classifications, and that over 75 percent of the students would be classified the same by the two criteria.

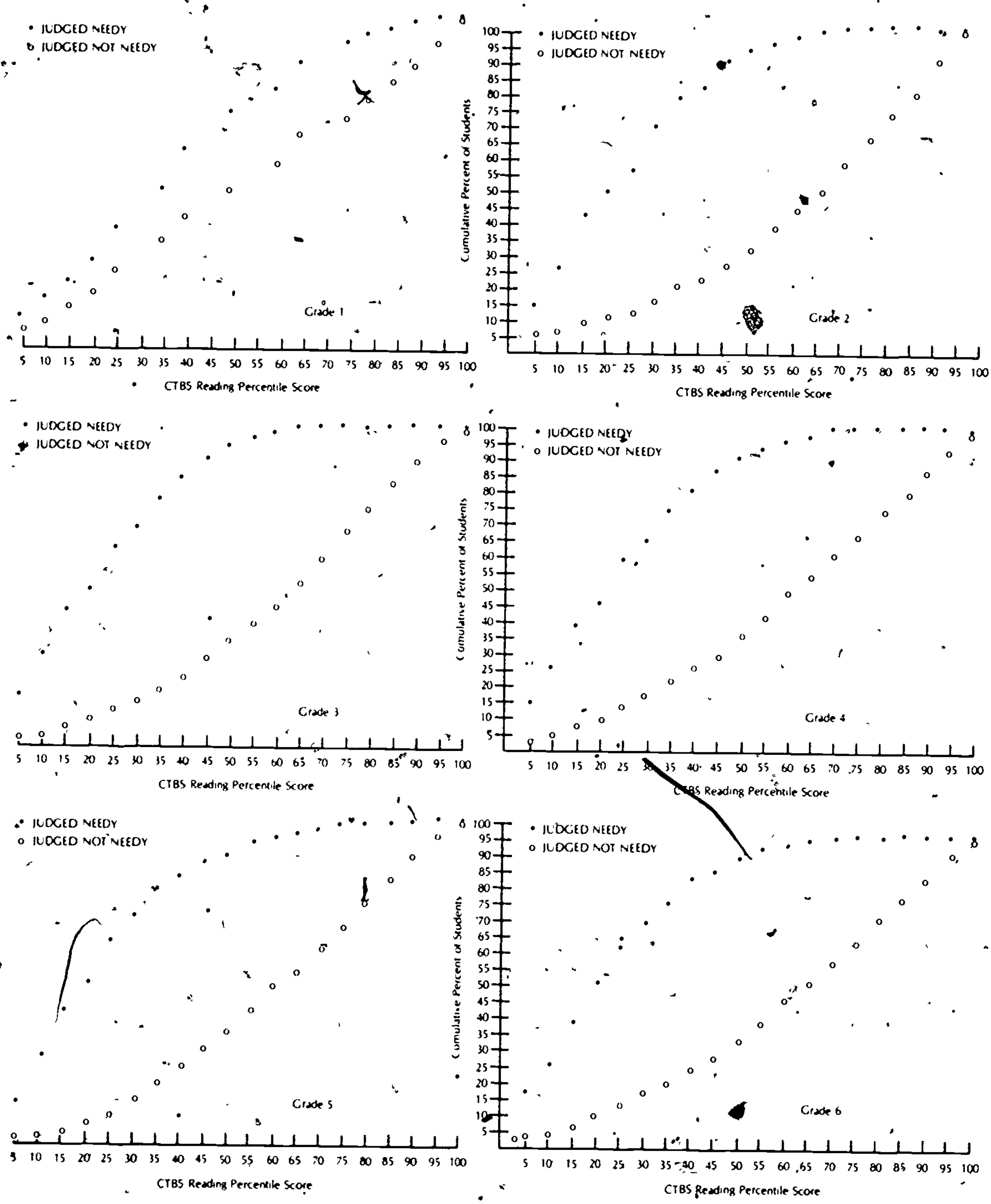

Figure 8-1

Cumulative Percentages of Reading Achievement for Students Judged To Need and Not To Need Reading CE

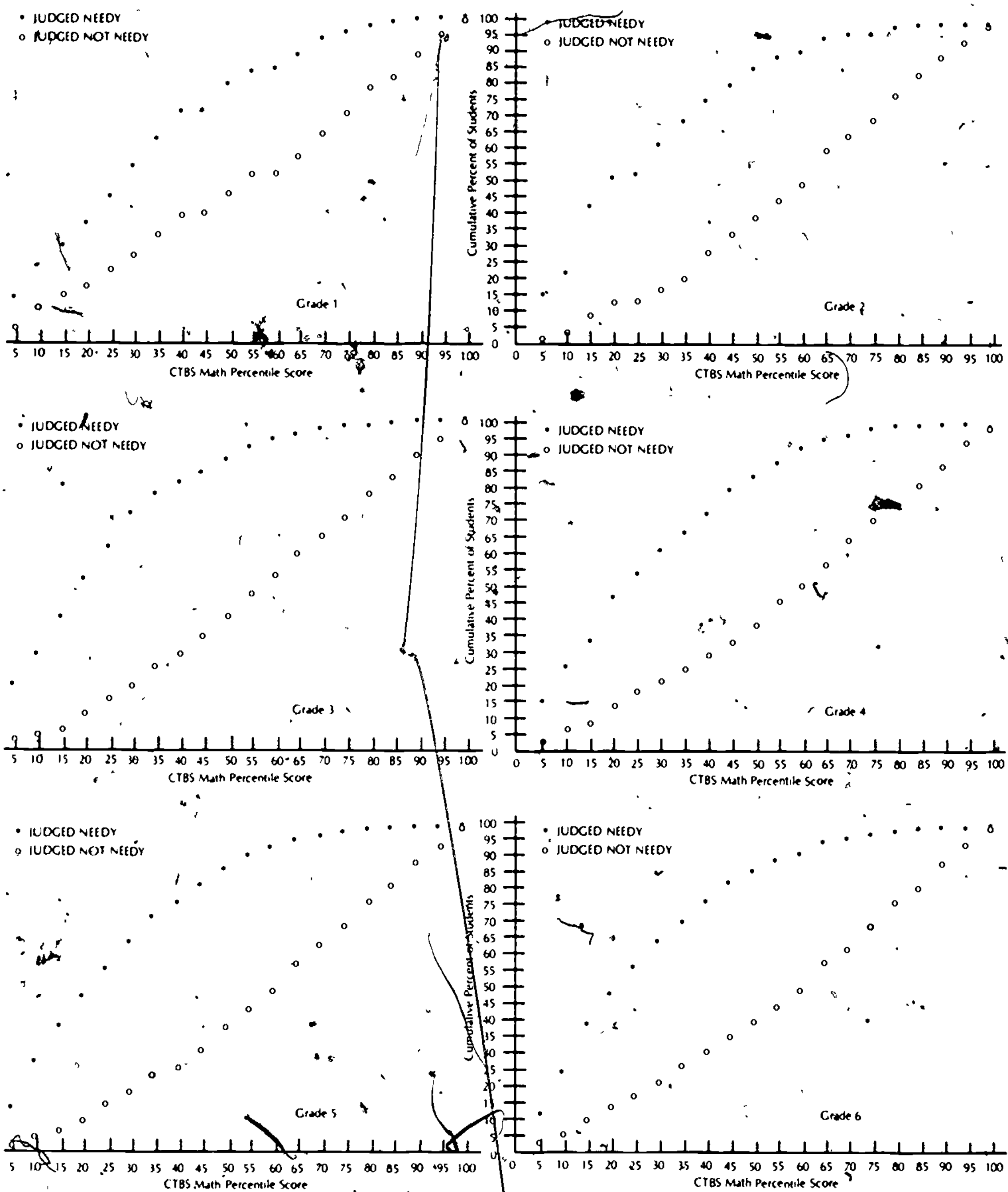

Figure 8-2

Cumulative Percentages of Math Achievement for Students Judged To Need and Not To Need Math CE

INFLUENCE OF OTHER STUDENT CHARACTERISTICS

Four student characteristics, all reported by the same teacher who judged need, were selected as potentially influencing teachers' judgements. student ethnicity (majority vs. minority), education levels of the student's parents (postsecondary education vs. no postsecondary education), student receipt or non-receipt of free or reduced-price lunch, and primary language spoken in the student's home. The impact of each of these variables was examined by employing it as a control variable in a partial correlation. (Actually, the partial correlation cannot be interpreted as an indicator of influence of the control variables on teacher judgment alone, since it includes their influence on both variables correlated, but it serves to explore the impact of the control variables on the correlation, as an initial step.)

For this analysis, both the average of correlations computed for each teacher and the correlations computed for the whole sample, by grade, were examined. The results were similar for both. Since the correlations at the whole-sample level were computed in the usual way while the teacher-level correlations were approximated, the whole-sample correlations were selected to report here. They appear in Table 8-4.

In every case, the partial correlation is somewhat smaller in magnitude than the simple correlation, indicating some influence of the control variables on the relationship between teacher judgment and achievement. The influence is uniformly small, however, the largest (for math in the third grade) being only .07. The simple correlations themselves account for no more than about 38 percent of the variance. The most accurate summary seems to be

Table 8-4

Simple and Partial Correlations Between Teacher Judgment and Achievement at the Whole-Sample Level, by Grade

Subject Correlation	Grade 1	2	3	4	5	6
Reading						
Simple	– 22	– 60	– 62	– 57	–.59	– 56
Partialing ethnicity out	– 20	– 57	–.57	– 54	– 55	– 51
Partialing parent ed out	– 18	– 56	– 56	– 51	–.53	–.51
Partialing free lunch out	– 19	–.56	– 57	–.51	–.55	– 53
Partialing language out	– 20	– 58	– 60	– 56	–.59	–.55
Math						
Simple	–.31	– 42	– 51	– 45	–.49	–.43
Partialing ethnicity out	– 26	–.38	–.45	– 42	–.45	–.38
Partialing parent ed out	– 26	– 38	– 45	–.39	– 42	– 39
Partialing free lunch out	–.25	–.38	– 44	– 40	– 45	–.37
Partialing language out	– 29	–.42	– 49	– 45	–.48	–.42

that teacher judgment tends to agree more than to disagree with achievement level, but neither variable is a potent predictor of the other (but compare the predictive power if a 'cutting score' is used to make the CTBS score a dichotomy). When the student variables are partialed out, the correlation is reduced slightly, indicating that student characteristics do not play a large role in the relationship between measured need for CE and judged need.

INFLUENCE OF TEACHER/SCHOOL CHARACTERISTICS

Several teacher or school characteristics may influence a teacher's judgements of student need for CE The first of these to be considered is the basis of student assignment to CE programs This information was obtained from an item of the Teacher Questionaire, Part B (for reading), and an item of Part C (for math). Both the items permitted multiple responses, and so could yield any combination of the following options:

- An achievement test score
- Teacher judgment
- Parent request
- Student request
- Other

For the present analysis, the responses were coded to yield three response categories as follows if 'teacher judgment' was marked, the response was assigned to the 'teacher judgment' category, regardless of other responses; if *only* 'an achievement test score' was marked, it was assigned to the 'achievement test' category, and if any other combination of marks was found, it was assigned to the 'other' category. This last category, then, includes any multiple response that does not have 'teacher judgment' as a component. The same coding was performed on both the reading and math items.

The three categories formed in this way, together with the six grade levels, defined the cells for a three-by-six analysis of variance. The basis of student assignment to reading CE was used in defining the cells for the analysis in which the Teacher Judgment Index for Reading (TJIR) was the dependent variable, and the corresponding item for math CE was used to define the cells when the Teacher Judgment Index for Math (TJIM) was the dependent variable Results of the analyses appear in Table 8-5 for TJIR and Table 8-6 for TJIM.

Confirming the previous pattern in the correlation means, the analysis of variance of the coefficients shows that TJIR and TJIM differ for different grade levels. This difference is to be expected on the basis of greater familiarity with the students, and is also influenced by the increase in validity of the CTBS after the first grade. But the primary variable for which this analysis was performed is the Basis of Assignment, it seemed reasonable that the usual practice of assigning students to CE could affect the teachers' perceptions of the importance of such judgments or their levels of experience in making such judgments. This school

Table 8-5

Analysis of Variance for TJIR

Source of Variation	Sum of Squares	df	Mean Square	F	Significance
Grade (G)	22 795	5	4 559	32.685	0 000
Basis of Assignment (B)	0 153	2	0 077	0 549	0 578
G X B Interaction	1 582	10	0.158	1 134	0.333
Residual	144 782	880	0.165		

Table 8-6

Analysis of Variance for TJIM

Source of Variation	Sum of Squares	df	Mean Square	F	Significance
Grade (G)	3 809	5	0.762	7.416	0.000
Basis of Assignment (B)	0 052	2	0.026	0.253	0.777
G X B Interaction	1 225	10	0.123	1 193	0.292
Residual	80.024	779	0.103		

characteristic has no significant effect on either TJIR or TJIM, either alone or in interaction with grade level. It appears, therefore, that a policy of making assignments to CE on the basis of teacher judgment has little or no effect on the quality of those judgments (as measured by their correlations with achievement scores).

Three composite measures of teacher characteristics or school practices appeared to be associated with accuracy of judgment, as indexed by TJIR and/or TJIM. One of these, Teacher Experience and Training, could equally logically be related to either index, and so correlations were computed for both. The other two composites are indices of the degree to which the instructional approach is individualized, and are expressed separately for reading and math, so only the correlations between the reading individualization index and TJIR, and between the math individualization index and TJIM, were computed. Correlations are presented separately by grade, in Table 8-7.

All but three of the correlations are less than .1, indicating that the variable being correlated with TJIR or TJIM would enable us to predict less than one hundredth of the variability in the index. The largest correlation, .134, accounts for only .018 of the variance. It is apparent, then, that none of these variables has any appreciable effect upon the accuracy of teacher judgment of a student's need for CE.

Table 8-7

Correlation of Teacher Judgment Index With Composites

Composite Measure	Grade 1	Grade 2	Grade 3	Grade 4	Grade 5	Grade 6	Whole Sample
Teacher experience and training							
With TJIR	- 04	- 00	- 09	13	- 07	01	01
With TJIM	- 00	01	-.13	05	-.03	.00	-.02
Individualized instructional approach							
Reading With TJIR	- 06	- 03	10	07	00	08	.03
Mathematics With TJIM	-.05	-.01	- 11	01	- 05	-.08	- 05

167

PART III. MEASURING STUDENT GROWTH

Part III contains four chapters on the selection and development of the measures of student growth and two chapters that investigate issues of testing that have been raised. In the first chapter we describe how the Comprehensive Test of Basic Skills was selected as the standardized achievement measures for the study, and how the tests were 'debiased' in an effort to make the scores apply equally well to all kinds of children. We describe the alternative to this kind of testing in the second chapter. The criterion-referenced approach was carefully studied to determine its applicability to the SES. In the third and fourth chapters, the selection and development of measures of functional literacy and computation and of affect are described in detail. Exhaustive searches for usable instruments were made first. In the case of the measure of practical achievement, no test was found that could meet the needs of the study, so a new test was developed, field-tested, and revised. In the area of student affect, a published instrument for measuring attitudes to reading and math was selected, but the instrument was augmented with new items that would bring the scores into closer agreement with the goals of the study.

In the fifth chapter we look at 'out-of-level' testing—a practice whereby very low-achieving students are tested with a level of test designed for students in a lower grade. Out-of-level testing has much to recommend it on several grounds, but it also presents problems. The pros and cons are discussed in some detail.

In the last chapter we investigate the issue of the 'speededness' of the achievement tests used in the SES. Our special concern is to determine if the speed factor influences scores differently for different racial/ethnic groups. In the event that such effects can be uncovered and speed is considered an irrelevant component of achievement, testing procedures in the SES should be modified to eliminate speed so that the findings are not influenced by it.

CHAPTER 9. MEASURES OF ACADEMIC GROWTH FOR THE SUSTAINING EFFECTS STUDY

Ralph Hoepfner and François Christen*

After careful evaluation of all available standardized achievement tests, the Comprehensive Test of Basic Skills (CTBS), Form S, was selected as the measure of growth in reading and math to be administered to all grades in both the fall and spring. The items of the CTBS were studied for sociocultural bias and those identified as biased were marked for elimination from total scores. Studies were also carried out to determine if alternate forms of the test should be used (they should not), if test levels should be counterbalanced in administration order during the first year (the order should be easier level first), and what kinds of scores should be prepared for later analyses (only reading and math total scores should be analyzed). The results of these efforts were a set of recommendations that would guide the achievement analyses for the remainder of the study.

Achievement growth in the areas of reading and math are almost universally considered to be the most important outcomes of our educational system. These skills, along with the affective and social skills that are receiving increased attention by educators, enable the individual to function effectively in today's complex society and to realize personal potential. For these reasons, the measures of academic growth for the Sustaining Effects Study (SES) are of critical importance.

SELECTING THE MEASURE OF ACADEMIC ACHIEVEMENT

There can be no doubt that published standardized achievement tests are the most widely used and the most exhaustively studied measures of academic growth. In addition, their historical popularity and acceptance have added to their perceived value as indicators of achievement growth. The task of selecting an achievenment test for a large national study of compensatory education (CE) cannot be taken lightly. The use of standardized achievement tests in educational evaluations has been criticized from many quarters. Carver (1974), for instance, found standardized achievement tests to be too psychometric as opposed to edumetric, claiming that they focus too much on stable between-individual differences rather than on within-individual growth. Others, such as Horst, Tallmadge, and Wood (1974) and Barker and Pelavin (1975), question the technical adequacies of achievement test as they are used for program evaluation.

The concern about using standardized achievement tests in educational research has also been mounted as a result of research findings on the impact of education on disadvantaged children. Most recent studies have shown that educational programs such as Head Start produce only small measured effects on educational achievement. As a consequence, researchers such as Campbell and Erlebacher (1970) have carefully scrutinized the

*Christen was affiliated with RMC Research Corporation, which was under subcontract to SDC to assist in selecting the achievement tests methodology used in past evaluation efforts. Had research found that education produced large measured effects on disadvantaged children, there probably would be less concern about the methodology used in the evaluation of educational programs and less concern over the adequacy of standardized achievement tests. The very refined statistical models now in vogue in educational research make increasingly greater demands on the psychometric properties of achievement tests.

Our review of existing achievement tests revealed that none had all the characteristics we desired for use in the SES Our strategy was, therefore, to select the instrument that best met our needs from those available.

The task of selecting the best instrument called for a trade-off analysis of standardized achievement batteries on the basis of the objectives of the study. An advisory panel on nationally known testing experts was to assist us in the development of selection criteria. Criteria were to be based upon a review of the study design, standards developed by the American Psychological Association, reviews conducted by the Center for the Study of Evaluation, and reviews sponsored by the federal government in conjunction with other recent evaluation studies. The selected test was to:

- Have empirically based fall and spring percentile norms
- Span grades 1 through 6
- Have reading (comprehension and vocabulary) and math (concepts and computation) subtest for most grade levels
- Assess basic skills rather that reflect a specific curriculum
- Have levels convertible to a common metric that cuts across grade levels
- Have minimum ethnic-group bias and be relevant, interesting, and meaningful to all students
- Have representation of minority groups in the standardization sample
- Have acceptable reliability and validity
- Have parallel forms
- Be easily administered, scored, and processed

A screening procedure was applied to the CSE Elementary School Test Evaluations (Hoepfner et al., 1976), a large compendium of tests evaluated by the Center for the Study of Evaluation (CSE) First, reading tests were screened using criteria developed by CSE for subject appropriateness, norm appropriateness, ease of scoring, and score interpretability. The surviving tests were then reviewed in terms of grade-level appropriateness. A parallel procedure was applied to math tests.

Reading Test Selection

Subject-Matter Appropriateness. The measurement requirements were first translated into content areas according to the CSE test evaluations. The following CSE educational goals were thought to be important in assessing reading proficiency:

Goal 1. Word Attack Skills. Identifies or combines the sound components (not the meaning) of words using phonetic skills or structural clues. Knows correspondence between sounds and their written representations. Identifies letters, syllables, roots prefixes, and suffixes.

Goal 2· Recognition of Word Meanings. Shows understanding of the meanings of written words, by identifing definitions, similar words, illustrations, synonyms, or antonyms. Knows different meanings that the same word may communicate.

Goal 3: Reading Comprehension. Understands material read. Infers the meaning of words from context Follows written directions. Identifies topic sentences, main ideas, and intentions of the author, and finds supporting details and illustrations in the text. Keeps track of temporal sequences, spatial order, and other relationships. Reads at a rate appropriate to the material and purpose.

Goal 1, above, is appropriate only for grades 1 and 2, and serves as a general 'reading readiness' category. Most tests in this goal category are word recognition (without meaning) or syllable-letter attack skills. Goals 2 and 3 are appropriate for grades 1 through 6 and even beyond Most tests in goal category 2 ask for the examinee to match words with something on the basis of the read meanings of the word. Tests in the third goal are of the 'read a paragraph and answer the question' type.

The first stage in the selection process consisted of eliminating those tests that do not have subscales assessing at least two of the above-mentioned goal categories (goals 1, 2, or 3 for grades 1 and 2; goals 2 and 3 for grades 3 through 6). On this basis, 58 reading tests were eliminated from further consideration.

Norm Appropriateness. The next criterion used for screening was norm appropriateness At this stage we again used the CSE evaluation, although the CSE criteria are not as rigorous as we desired. The CSE criterion for norm appropriateness is as follows:

> Is the norm group representative of the national population? Norms should be based on a sample of examinees that is drawn from a variety of family and community backgrounds and includes students at the grade level for which the test is being evaluated. Raters looked for explicit or implicit evidence that the sampling was intended to be random or stratified, recent, geographically balanced, racially and ethnically representative, and heterogeneous with respect to population density. The test or subscale was credited with 1 point if an attempt was clearly made to obtain a nationally representative and current sample. It was given no points if there was no norm sample, or the sample appeared to be local or incidental (Hoepfner et al., 1976).

Forty-three of the remaining tests did not receive a rating of 1 on the CSE evaluation and were consequently eliminated.

Ease of Scoring The reading tests were then screened for ease of scoring. In a large study it is imperative to have a machine-scorable test The CSE test evaluations do not cover the issue of machine scorability, but do rate tests on ease of scoring. The CSE criterion is as follows.

> How easy and objective is the scoring procedure? Hand or machine scoring that can be quickly carried out is, of course, preferable to more difficult and time-consuming procedures. Scoring was judged to be 'objective and simple' whenever it consisted of some straight-forward process. Tests fulfilled this condition in many ways, providing answer sheets, matching stencils, templates, machine-scoring services, or just uncomplicated scoring guides

The test or sub-scale was credited with 2 points if the scoring procedure was objective and simple Only one of the remaining tests did not earn 2 points, it was eliminated.

Score Interpretability The next criterion on which tests were screened was score interpretability Here again, we used the CSE test evaluations to eliminate test with unusual or naive score conversions The more technical issues relating to grade-equivalent scores and to a score metric common over all grades, however, are discussed in greater detail later. The CSE criterion was as follows:

> Can the scores be easily interpreted relative to some norm group or standard? Converted scores that are common, simple and clearly explained are desirable in order that school personnel will understand the scores and not be misled by them Pass/fail, percentile ranks, stanines, and grade equivalents are considered common and simple. The test or subscale was credited with 1 point if common, simply interpreted conversions of test scores were provided, and 0 points if no conversion from raw scores is provided, or if conversions are novel, less common, or unclear in their meaning.

One of the remaining tests did not meet this criterion and was eliminated.

Grade-Level Appropriateness. The next screen was employed to eliminate tests that did not cover grades 1 through 6. Six tests were eliminated.

Surviving Reading Tests. The following eight tests met the broad and permissive screening criteria for the measurement of reading goals:

California Achievement Test
Comprehensive Test of Basic Skills
Durrell Listening-Reading Series
Gates-MacGinitie Reading Test
IOWA Tests of Basic Skills
Metropolitan Achievement Test
SRA Assessment Survey
Stanford Achievement Test

Math Test Selection

The procedure used to screen math tests was parallel to the one adopted for reading tests. However, since there are fewer specialized math tests and since the CSE rules for classifying a test in one goal category as opposed to another are perhaps more arbitrary for math than for reading, tests were not eliminated as to goal appropriateness on the first round unless they covered none of the goal categories outlined below.

Subject Matter Appropriateness. The following CSE educational goals were thought to be important in assessing proficiency in math:

Goal 1 Knowledge of Numbers and Sets. Understands numbers and fractions. Differentiates between numerals and between prime and composite numbers. Identifies factors, multiples, and relative primes of a given number. Understands set membership, set relations, set correspondence, and operations with sets. Relates set notation and diagrams to categorical statements in English.

Goal 2 Knowledge of Numeral Systems and Number Principles. Reads, recites, and writes numerals. Understands place values, the rounding of numbers, the decimal system of numeration, numeration with bases other than 10, and Roman numerals. Understands the commutative, associative, and distributive properties, inverse operations, properties of 0 and 1, negatives, and reciprocals. Understands number-line diagrams. Finds and evaluates simple numerical rules based on observation.

Goal 3 Knowledge Basic to Algebra. Understands number relationships, number sentences, variables, and formulas. Reads sentences using letters or frames and equality or inequality signs, and relates them to quantitative statements in English. Solves or graphs equations and inequalities. Tests relations for reflexivity, symmetry, and transitivity.

Goal 4 Whole-Number Computation. Adds, subtracts, multiplies, and divides integers, checks answers.

Goal 5 Decimal and Percentage Computation. Adds, subtracts, multiplies, and divides decimals, or decimals and integers, checks answers. Transforms fractions into decimals, decimals into percents, percents into fractions, and vice versa. Reads and writes decimals and percents. Solves percentage computation problems.

Sixty-nine tests with sub-scales that assess at least one category from goals 1 through 3 (concepts) or from goals 4 through 6 (computation) were included for further screening. Tests that fall into goal categories 1, 2, or 3 tend to have items in other goals also, although CSE's procedure arbitrarily categorizes them into one goal rather than another. Tests that cover goals 1, 2, or 3 are appropriate for grades 1 through 6 and beyond. Tests of goal 4 are appropriate for all grades, tests of goal 5 for grades 4 and above, and tests of goal 6 for grades 6 and above

Norm Appropriateness Math tests that do not have appropriate norms according to the CSE test evaluations were eliminated. The CSE criterion is the same as was previously defined for the reading goals. Fifty-four tests did not meet it.

Ease of Scoring. The next screen of math tests eliminated those that did not meet the CSE test evaluation's criterion of ease-of-scoring, as previously defined. Only two tests were eliminated because they did not meet this criterion.

Score Interpretability. The remaining tests were subjected to CSE's criterion of score interpretability as previously defined. Only one test did not meet the criterion.

Grade-Level Appropriateness. Six math tests did not cover grades one through six and were therefore eliminated.

Surviving Math Tests. The following tests successfully fulfilled the general criteria for math tests for use in the SES:

California Achievement Test
Comprehensive Tests of Basic Skills
IOWA Tests of Basic Skills
Metropolitan Achievement Test
SRA Assessment Survey
Stanford Achievement Test

Because the SES is focused on educationally disadvantaged children, it was expected that there would be a substantial amount of out-of-level testing—that is, children would be given a test lower than that recommended by the test publisher for a given grade (see Hoepfner and Wang, this volume). Out-of-level testing is used to ensure that the test given is not so difficult for the students that a meaningful measure of their academic achievement cannot be obtained. Consequently, it is necessary to choose a test of academic achievement that covers one grade level below the lowest grade level being tested. Since the lowest grade level tested in the SES is grade 1, the academic achievement test selected should cover kindergarten. Thus, for both reading and math, four of the surviving tests were eliminated because they did not cover kindergarten. The remaining tests under consideration were:

Comprehensive Tests of Basic Skills
Gates-MacGinitie Reading Tests
Metropolitan Achievement Test
Stanford Achievement Test

Additional Selection Criteria and Visits to Test Publishers

The expert panel on achievement tests made several recommendations for test selection. The panel first recommended that of the four test finalists two be seriously considered and scrutinized The Comprehensive Tests of Basic Skills (CTBS) and the Stanford Achievement Test (SAT). The publishers of these two tests were visited to obtain both technical and marketing information about the tests. The technical and marketing information sought, as well as a report of the findings from the visits to the test publishers, are discussed below.

Because there will be out-of-level testing in the SES, it will be important to be able to convert a student's score on one level of the test to a scale that covers all levels, in order to compare students in a given grade taking different levels of the test. Vertical scales are provided by most test publishers to do just this. However, studies indicated that vertical scales were inadequate when used with disadvantaged children. It was thus imperative to determine how the test publishers derived the vertical scales for the two tests under consideration, and whether they made any empirical checks on its adequacy.

Not all achievement test publishers provide fall and spring norms, based on empirical data. Many interpolate norms for periods where no actual data were collected. Projected norms are generally based on the assumption of linear cognitive growth over the calendar year. Since there is no evidence to support the linear-growth assumption, the created norms may be inaccurate enough to give a distorted view on the impact of CE. The panel on achievement tests recommended that, other things being equal, a test should be chosen that offers empirical fall and spring norms. It had been suggested by the panel that the SES could in itself provide the data necessary to generate better norms than the publishers could supply. A second consideration related to a test's norms is the quality and recency of the norm sample If scales based upon the norm sample were to be employed in this study, the quality of the sample might be a critical consideration.

Statistical analyses are increasingly being used to detect whether achievement tests are biased. In essence, a two-way analysis of variance is carried out, where race is one factor and test item is the other factor. If an interaction between race and test item is found, at least one component item is said to be biased. The test is made unbiased by removing test items that interact with race (Ozenne, Van Gelder, and Cohen, 1974). To perform such an analysis, it is important that the publisher have individual item data for students of different race/ethnicity.

In choosing the achievement test, the following additional information had to be determined from the test publisher. (1) whether the test could be rapidly scored by a process that reads character labels, (2) the royalty rates that applied to test use and scoring, (3) the possibility of reprinting selected subtests of the test in a format more appropriate to the study, (4) the restrictions involved, and permissions needed for reprinting, and (5) the royalty rates that applied to the test printing.

The visits to the test publishers clarified and revealed the following additional information regarding the two contending test series:

- Both publishers were eager to cooperate in making both data and computer programs available

- Both publishers felt that [illegible]y had good normative data. Only the SAT had empirical fall and spring norms at [illegible]e levels, the CTBS had empirical fall and spring norms only at the lower levels.

- Both publishers agreed that new empirical norms could be created by re-weighting the study sample Indeed, they encouraged this procedure.

- Both publishers use a modification of Thurstone's method of absolute scaling to calibrate various test levels onto a common metric. However, two levels of the test are never administered to the same group of children in the calibration process. In order to effect the calibration procedure, test publishers made additional statistical assumptions which may be unwarranted.

- Vertical scale scores are not available for all the sub-tests that would be chosen for the study. Thus, they must be created either by the publisher or from data collected during the course of the study.

- Only the CTBS publisher collected ethnic data from the standardization sample. The availability of such data would allow for carrying out a de-biasing study at the outset of the study rather than waiting for the first-year results.

- Both publishers provided practice tests. These were used with the standardization sample.

- Both publishers expressed willingness to create new norms based on the subscales chosen.

Final Test Selection

After considerable scrutiny, it was evident that the SAT and the CTBS were equally good representatives of the contemporary achievement tests, however, both had some shortcomings for the SES. First of all, as complete batteries they take too long to administer. Second, the technical property deemed most important for their use in a longitudinal study—the vertical scale—had not been empirically validated or verified. Third, norms were not based on a national probability sample. Thus, the SES had to overcome the shortcomings of the instruments. Consequently, the selected test would not be used as a complete instrument for the measurement of academic achievement, but rather, as a basic resource in the selection of subscales that cover reading and math. The recommendation then was that the final selection should be based not on technical grounds, but rather on pragmatic considerations.

The CTBS was chosen for the SES for the following pragmatic reasons:

- It would be possible to carry out a de-biasing study based on the publisher's data before the first wave of data collection, and thus to create revised scoring keys for immediate use with the first-year data.

- The item layout appeared to be better for the lower levels of the CTBS than for the lower levels (Stanford Early School Achievement Test, SESAT) of the SAT.

- The CTBS has been more recently revised than the SAT and thus probably reflected current pedagogical approaches better.

- The CTBS publisher had computer software compatible with that of SDC.

- The CTBS had been subjected to one round of statistical de-biasing, whereas the SAT had not.

When the CTBS was selected as the standardized achievement measure, several additional problems had to be resolved in order to implement the test in an effective and responsible manner: identification and elimination of biased items; the use of alternate forms; order and learning effects in administering two levels of the test; and level of scores to be analyzed. Two other critical problems, the development of improved norms and vertical-scales and the effects of test speededness are discussed in detail in Report 10 and in another paper on the present report, respectively.

IDENTIFICATION AND ELIMINATION OF BIASED ITEMS

The use of standardized achievement tests in previous evaluations of CE has caused numerous educators, evaluators, test developers, and minority group members to express concern that such instruments are not appropriate as measures of academic performance for minority students. In the main, criticisms of standardized achievement tests have focused on the claim that such tests were developed for Anglo, middle-class students and are biased against minority students.

The Origin, Nature, and Consequences of Sociocultural Bias in Tests. A major concern in the SES was focused on sociocultural bias in tests. That is, we were particularly concerned with biases resulting from systematic differences in the sociocultural backgrounds and experiences of members of certain groups, where these backgrounds and experiences differ substantially from those of other groups. More specifically, we were concerned with bias against disadvantaged and minority groups to the extent that the sociocultural history, environment, and values of those groups are known to differ from those of the children on whom the tests are normed. The consequences of sociocultural test bias depend on how the test data are used. All too often, data from biased tests are improperly used to infer underlying, innate differences between different racial or cultural groups. Such an interpretation may be used to justify 'giving up' on certain segments of the population, i.e., to assume that it is pointless to make any effort to change things that purportedly have inherent or hereditary origins. Bias in tests may also influence decisions about who should be promoted in school, assigned to advanced learning groups, given other signs of recognition such as membership in clubs and societies, accepted into colleges, and hired in desirable jobs. Use of biased tests may produce systematic underrepresentation of disadvantaged and minority students in the benefit of these forms of recognition.

Finally, and most relevant to the SES, use of a biased test may lead one seriously to underestimate the value of a CE program such as Title I. Such a test could underestimate, for disadvantaged and minority students, not only the absolute achievement but also the gain in achievement that is used as a criterion of program success.

Given these concerns, we conducted a study of the potential bias in the CTBS. Results of the examination of the CTBS for biased items would allow us to perform subsequent analyses with some certainty that data will not be overly influenced by test items that seem culturally biased against particular ethnic groups. Review of the CTBS for bias was accomplished in two phases. Briefly, the first phase involved the identification of items that are statistically biased* against ethnic groups as demonstrated by the publisher's standardization data. In the second phase, a ten-member panel of persons from a range of ethnic groups evaluated the statistically biased items to determine which contained content that was culturally biased against certain ethnic groups. The panel members' review of item content was carried out by using a modified Delphi approach, with two cycles of item review.

Statistical Analyses to Identify Biased Items. The statistical procedures employed in identifying potentially biased items were based upon the notion that items that appear to 'work differently' for different groups are likely to be biased. Several approaches to the operationalization of 'work differently' have been utilized in previous studies attempting to determine item bias. Most recently, Ozenne, Van Gelder, and Cohen (1974) utilized two statistical-analysis approaches, coupled with professional judgment, in the identification of item bias in the California Achievement Tests. Their approaches are described below.

The first statistical approach can be labeled an 'analysis-of-variance' approach to the identification of potentially biased items. This approach is based upon the work of Cleary and Hilton (1968). Item response data are analyzed within a two-way factorial model for analysis of variance, with group membership (ethnicity or race) one of the ways, and items the other way. In this type of analysis of variance, there will be two main effects and one interaction. Evaluation of the main effects would indicate whether the ethnic groups are different in overall test level and whether the items are different in difficulty. The main effects are, for this type of analysis, overlooked in favor of the interaction effect. A significant interaction indicates that some (at least one) items are working differently from the way most others work. The interaction effect takes into account that the ethnic groups may be different in overall level and that the items may have different difficulty levels, and then looks for items that are working in ways not expected from these main effects.

Because this analysis-of-variance design is a repeated-measure design, and because the number of items analyzed for any one test or subtest is relatively large, a rigorous statistical interpretaion of any significant interaction is difficult to make. The problems, of course, are in accounting for the correlated nature of the item responses (the same students answer all the items) and then identifying, through some sort of after-the-fact tests, precisely which items are causing the significant findings. The rigorous statistical identification of items was avoided by Ozenne et al., by substituting graphic methods for detecting those items that were 'working' differently.

*It should be noted that the correspondence between statistical bias and bias as usually meant in terms of justice or morality is based on a number of inferences that are not compelling to all researchers who study the problem.

For the SES, a two-factor design with repeated measures on one factor (Winer, 1971) was employed in the analysis of each skill area and test level. Subtests within skill areas were analyzed jointly, as the total skill-area scores were to be employed in the study, and it is these total scores that should be free from bias. Analysis of each of the subtests separately might fail to identify items that are biased in part because of the bias of the subtest of which they are part For each test level, one grade of students was selected. Selection of the one grade per level that provided the largest number of students 'sharpens' the analyses by not allowing between-grade variances to enter into the comparisons. Because in the publisher's norm sample the ethnic grouping of brown students was always the smallest in number and because the computer program (BMD 08V) calls for an equal number of cases in each cell, the grade level for the analysis of each test level was selected to maximize the number of brown students. The number of brown students was then matched by randomly selecting equal numbers of black and other students. In this way, sample students from each ethnic group were selected for the following test levels. 222 from kindergarten for Level A; 318 from grade 1 for Level B, 231 from grade 2 for Level C; 382 from grade 3 for Level 1 457 from grade 5 for Level 2, 999 from grade 7 for Level 3, and 717 from grade 9 for Level 4 Summary tables from the two-factor, repeated measures analyses of variance are provided in Table 9-1

Inspection of Table 9-1 indicates that in each of 14 analyses, both of the main effects and the interaction are significant The interpretation of the 'ethnicity' main effect, as provided by the overall means, is that in all cases, the black students earned the lowest mean score and the other students earned the highest mean score. (One must, of course, guard against any obvious interpretation of these consistent findings, as ethnicity in this sample, is undoubtedly confounded with a host of socioeconomic and cultural variables.) The significance of the 'item' main effect, indicates that the items have a wide range of difficulty levels In general, within a separately timed subtest, the early items are relatively easy, while the late items are relatively difficult The significant interaction terms indicate that some items are differentially difficult over the three groups. It is these items that were identified as statistically biased.

The method employed to determine specifically which of the items are so biased is a variant of the method Winer (1971, pp. 529-531) called a test of simple main effects. In the variant method employed, however, the main effects are ignored and the simple interaction effect is studied This is accomplished by analyzing each item separately over the three groups, as in a simple 1 x 3 analysis of variance. The difference is that the overall group-effect mean is subtracted from each subject's responses within each group, thus eliminating the main effect of ethnicity, and the error term, the denominator of the test, is an estimate of the pooled-within sum of squares, taken as an average over all the items. The error estimate is then employed as the denominator for each of the item simple-interaction-effect tests The resulting F ratio does not have an F distribution, but its distribution may be approximated by another F distribution with different degrees of freedom (Winer, 1971, pp. 530-531) equal to one less than the number of groups and (as a conservative-test estimate) the product of the number of groups and one less than the number of replications. The conservative estimate was selected because of its simplicity and because the large sample (number of replications) provided for degrees of freedom greater than those typically tabulated anyway.

Table 9-1

Repeated-Measures Analyses of Variance for Seven Levels of CTBS Skill-Area Scores

Test Level	Source	Reading Skills Area df	Reading Skills Area Mean Square	Reading Skills Area F	Math Skills Area df	Math Skills Area Mean Square	Math Skills Area F
A	Ethnicity	2	98 52	52 473*	2	35 65	42 281*
	Items	48	10 06	61 501*	25	20 87	115 351*
	Replications nested in ethnicity	663	1 88		663	0 84	
	Ethnicity x item interaction	96	0 34	2 088*	50	0 67	3 698*
	Replications x items nested in ethnicity	31 824	0 16		16 575	0 18	
B	Ethnicity	2	177 31	75 732*	2	157 45	82 753*
	Items	42	9 55	55 401*	45	11 93	62 370*
	Replications nested in ethnicity	951	2 34		951	1 90	
	Ethnicity x item interaction	84	0 32	1 850*	90	0 62	3 242*
	Replications x items nested in ethnicity	39 942	0 17		42 795	0 19	
C	Ethnicity	2	96 24	34 868*	2	39 63	21 509*
	Items	50	12 21	81 644*	39	11 82	66 841*
	Replications nested in ethnicity	690	2 76		690	1 84	
	Ethnicity x item interaction	100	0 29	1 934*	78	0 27	1 555*
	Replications x items nested in ethnicity	34 500	0 15		26 910	0 18	
1	Ethnicity	2	467 41	132 707*	2	346 10	115 351*
	Items	84	20 21	111 141*	72	24 26	143 141*
	Replications nested in ethnicity	1 143	3 52		1 143	3 00	
	Ethnicity x item interaction	168	0 42	2 324*	144	0 59	3 484*
	Replications x items nested in ethnicity	96 012	0 18		82 296	0 17	
2	Ethnicity	2	433 40	132 416*	2	192 89	82 079*
	Items	84	28 26	154 128*	72	44 49	251 440*
	Replications nested in ethnicity	1 368	3 27		1 368	2 35	
	Ethnicity x item interaction	168	0 58	3 153*	144	0 61	3 461*
	Replications x items nested in ethnicity	144 912	0 18		98 496	0 18	
3	Ethnicity	2	1 371 55	446 667*	2	906 31	350 846*
	Items	84	59 67	326 244*	72	53 69	285 232*
	Replications nested in ethnicity	2 994	3 07		2 994	2 58	
	Ethnicity x item interaction	168	1 60	8 75[illegible]	144	1 51	8 011*
	Replications x items nested in ethnicity	251 496	0 18		215 568	0 19	
4	Ethnicity	2	809 54	319 843*	2	751 73	276 635*
	Items	84	32 80	169 981*	72	50 06	278 324*
	Replications nested in ethnicity	2 148	2 53		2 148	2 72	
	Ethnicity x item interaction	168	1 46	576*	144	1 51	8 419*
	Replications x items nested in ethnicity	180 432	0 19		154 656	0 18	

*Significant at the .01 level

150

These tests were systematically applied to each item within each skill area at each level. Although the ideal method for the evaluation of the resultant single-factor test would be to predetermine the acceptable Type II error (accepting the hypothesis that a truly biased item is not biased), the sample sizes within cells were sufficiently large that the more convenient method of pre-setting an *alpha* would not likely cause any differences in the identification of potentially biased items These analyses, *alpha* was conservatively set at .001, partly because of the questionable assumptions made on the covariance matrices and partly because the resulting F statistic does not truly have an F distribution. Sixty-seven items with significant F tests are listed in Table 9-2, with their respective subtest codes and the original item numbers from Form S of the CTBS.

It can be seen from Table 9-2 that, as in the study by Ozenne et al. (1974), there are fewer items identified at the lower levels than at the higher levels, and not all the items appear to be biased against the minority groups. Among the 36 identified reading items, 4 appear biased against black students, 8 against brown students, 12 against other students, 10 against both black and brown students, and 2 against brown and other students. Among the 31 identified math items, 12 appear biased against black students, 2 against brown students 14 against other students, 2 against black and brown students, and 1 against black and other students Inspection of Table 9-2 will also reveal that there is not a consistent trend for early (easy) or late (hard) items to be biased against any one or two groups. If the analysis-of-variance approach to the identification of biased items is accepted as a meaningful way to study test bias, then these results indicate that the CTBS is fairly evenly biased against all three groups

The second statistical approach can be labeled a 'correlational' approach to the identification of potentially biased items This approach has been utilized by several test publishers in their efforts to eliminate items giving the appearance of being biased. In this approach, items are analyzed through the correlation of item-response pattern to total-score pattern within each racial/ethnic group, so that items having different correlational levels among the groups can be assumed to be measuring different things via the notion of internal-consistency reliability Indeed, this correlational approach is equivalent to the examination of differences among the item-total biserial correlation coefficients for each of the racial/ethnic groups The biserial coefficient should be employed because of its relative resistance to effects of the general difficulty level upon its size (Oosterhof, 1976). Because of difficulties in the estimation of the standard errors of differences between biserial coefficients however the analyses were performed upon point-biserial coefficients. Each triplet of point-biserials can then be tested for each item to determine if two of the coefficients are statistically different across the racial/ethnic groups (the intercorrelation among test items, the repeated-measures aspect of the first type of analyses, cannot be considered in these simple univariate tests) Ozenne et al. again utilized a judgmental approach to the identification of those items that seemed to have different item-total correlations for their different groups

In the SES, point-biserial coefficients between items and their totals were computed for each racial/ethnic group for all subtest within skill areas. The method for determining whether or not the coefficients were different was a simplified test based upon the point-

Table 9-2

CTBS Items Identified as Biased by the Analysis of Variance Approach, and Group(s) Biased Against

Test Level	Skill Area	Subtest Name	Item Number	Group(s) Biased Against
B	Math	Computation	32	Other
1	Math	Computation	1	Other
2	Reading	Comprehension	9	Black Brown
2	Math	Concepts	6	Brown
2	Math	Computation	8	Black Brown
2	Math	Computation	45	Other
3	Reading	Vocabulary	2	Brown Other
3	Reading	Vocabulary	3	Black
3	Reading	Vocabulary	6	Black Brown
3	Reading	Vocabulary	8	Black Brown
3	Reading	Vocabulary	10	Black Brown
3	Reading	Vocabulary	14	Other
3	Reading	Vocabulary	21	Black Brown
3	Reading	Vocabulary	22	Brown
3	Reading	Vocabulary	25	Black Brown
3	Reading	Vocabulary	28	Brown
3	Reading	Vocabulary	35	Other
3	Reading	Vocabulary	37	Other
3	Reading	Vocabulary	39	Brown
3	Reading	Comprehension	1	Other
3	Reading	Comprehension	6	Other
3	Reading	Comprehension	11	Other
3	Reading	Comprehension	13	Brown
3	Reading	Comprehension	22	Black
3	Reading	Comprehension	38	Other
3	Reading	Comprehension	41	Other
3	Reading	Comprehension	43	Black Brown
3	Math	Concepts	2	Black
3	Math	Concepts	7	Brown
3	Math	Concepts	8	Black Brown
3	Math	Concepts	14	Other
3	Math	Computation	1	Other
3	Math	Computation	10	Black
3	Math	Computation	13	Other
3	Math	Computation	23	Other
3	Math	Computation	26	Other
3	Math	Computation	31	Black
3	Math	Computation	32	Black
3	Math	Computation	36	Black Other
4	Reading	Vocabulary	1	Brown
4	Reading	Vocabulary	2	Black
4	Reading	Vocabulary	8	Black
4	Reading	Vocabulary	12	Black Brown
4	Reading	Vocabulary	13	Brown
4	Reading	Vocabulary	15	Black Brown
4	Reading	Vocabulary	19	Black Brown
4	Reading	Vocabulary	20	Brown
4	Reading	Vocabulary	25	Other
4	Reading	Vocabulary	27	Brown
4	Reading	Vocabulary	38	Other
4	Reading	Vocabulary	38	Other
4	Reading	Comprehension	6	Other
4	Reading	Comprehension	22	Brown Other
4	Math	Concepts	1	Black
4	Math	Concepts	4	Black
4	Math	Concepts	8	Other
4	Math	Concepts	10	Other
4	Math	Concepts	16	Other
4	Math	Concepts	23	Other
4	Math	Computation	3	Black
4	Math	Computation	5	Black
4	Math	Computation	6	Other
4	Math	Computation	8	Other
4	Math	Computation	16	Black
4	Math	Computation	17	Black
4	Math	Computation	27	Black
4	Math	Computation	31	Black

biserial's essential identity to the Pearson r (a characteristic not shared by the biserial coefficient), and thereby to its sampling characteristics. The point-biserials were tested in the following manner the largest and smallest coefficients for each of the 887 items were transformed into Fisher z's, and then a z test was made of their differences (Guilford, 1965). Because of the large sample sizes involved in these tests, *alpha* was set at .001.

Of the 887 triplets of coefficients thus tested, 155 exhibited significant differences. Table 9-3 presents the 155 item numbers, according to test level, skill area, and subtest, that were identified as biased by this correlational method. The results of the correlation analyses are similar to those of the analyses of variance, as they both indicate more biased items at the higher levels than at the lower levels These data are unable to address whether this

Table 9-3

CTBS Items Identified as Biased by the Correlational Approach

Level	Skill Area	Subtest Name	Item Number(s)
A	Reading	Sound Matching	21 26
A	Reading	Letter Sounds	5, 14
A	Math	Mathematics	13
B	Reading	Comprehension	16, 21
C	Reading	Comprehension	1
1	Reading	Vocabulary	13 22 27 31 33 39
1	Reading	Comprehension	12 15 22 24 38, 39 40 43
1	Math	Concepts	5 19 23
1	Math	Computation	20 23 24 30
2	Reading	Vocabulary	4 21 24 28 32
2	Reading	Comprehension	9 37
2	Math	Concepts	3 6 24
2	Math	Computation	2 5 9 26 32 42
3	Reading	Vocabulary	2 3 10 15 17 19 20 21 23 24 26 28, 29 30 31 34 37 39 40
3	Reading	Comprehension	2 4 9 16 29 31 38 39 43 44 45
3	Math	Concepts	4 5 16 18 19 20 21 22 24 25
3	Math	Computation	4, 5 6 7 8 11 12 15 17 18 19 20 21 22 29 30 32, 38 39 41 43 44 46 47
4	Reading	Vocabulary	8, 12 17 19 24 27 29 30 34 35
4	Reading	Comprehension	6 19 34 40 43 45
4	Math	Concepts	2 3 9 10 11 17 19 25
4	Math	Computation	1 4 9 10 11 14 15 17 20, 21 22 23 24 30 31 32, 36 41 42 44 46 47

phenomenon is due to large samples at the higher levels, or to some sort of within-group homogeneity and between-group heterogeneity that develops with age.

All the items that were associated with significance, either in the analysis-of-variance or the point-biserial examinations, were considered 'suspicious' items that would undergo judgmental analysis. A total of 201 items (67 identified by analysis of variance, plus 155 identified by correlational analysis, minus 21 that were identified both ways) were submitted to the bias panel, along with 14 additional items (one each for reading and math at each of seven test levels) that were identified as being statistically least biased (minimum observed differences in the statistical analyses). The unbiased items were included in order to support a caution to the judges that not all items had been identified as biased.

Judgment Analyses to Identify Biased Items. The items identified by the above procedures were suspected of bias because of their differing statistical properties in different subgroups. However, the statistical analysis does not indicate exactly what property of an item makes it biased. In fact, the statistical analysis may identify many items that are not biased but that have aberrant statistical characteristics only. Since the objective is to eliminate items that are truly biased against any of the subgroups involved in the study, it was necessary to identify the source of potential bias in each item before removing it from its test. A review of all the statistically biased items was therefore conducted to determine whether each item had a content or form that could bias it against one or more of the ethnic-cultural subgroups participating in the SES.

To evaluate the items determined to be statistically biased, a panel of ten people from a variety of ethnic backgrounds was used. *Asian Panelists.* From Hawaii, an elementary administrator and district Title I Director, from California, an urban elementary teacher. *Black Panelists.* From Mississippi, an expert in item construction and test development, and director of testing at a predominatly black university, from California, an experienced metropolitan elementary teacher, from South Carolina, an experienced urban elementary teacher and Title I coordinator. *Chicano Panelists.* From Texas, an expert in item construction and test development who had experience teaching Chicano students, from California, a community leader. *Indian Panelist.* From New Mexico, an expert in item construction and Head Start Teacher of Indian Students. *White Panelists.* From West Virginia, an experienced elementary and special education teacher of Appalachian students, from Massachusetts, an educational desegregation specialist.

The panelists carried out their review of the content of statistically biased items following a modified Delphi approach. Two separate reviews of items were made. In the first review, panelists were sent copies of all levels of the CTBS along with instructions for their review. In each of the test booklets, items to be reviewed were marked (and, for the lower levels, directions to be read by the test administrator were included), and the instructions indicated that the panelist was to rate each item as either 'biased,' 'possibly biased,' or 'not biased' on a rating form. No item was rated 'biased' or 'possibly biased' by more than six panelists. On the other hand, 99 of the 215 items were so rated by at least one panelist. A conservative criterion was used in determining which items were to be considered again during the second review, that of including all items rated 'biased' or 'possibly biased' by at least two panel members. Thirty-two items met this criterion.

For the second review, panelists were sent forms for each of the 32 items, along with comments made by panelists in the first review, and were asked to rate the items as either 'biased' or 'not biased.' Panelists were instructed to note which comment they subscribed to, and to include additional comments for any item that was rated 'biased.'

In order for an item finally to be characterized as being biased, criteria were established that at least five of the panelists had to rate the item as biased in the second review, and the item had to be statistically biased (the latter is important, since two of the 'control' items were rated biased by five or more panelists). These criteria resulted in the identification of seven items as culturally biased. One additional item was determined to be biased which did not meet the above criteria, it is an item that was statistically biased against brown students and the two Chicano panel members indicated that the key word used in the item stem had a Spanish root that would apply equally to two of the response alternatives. Thus, a total of eight items were identified as being both statistically and culturally biased.

All of the eight biased items come from reading subtests, none of the math subtest items were judged to be biased. Grouped by test level, characteristics of the biased items are indicated below.

Level A Two items, from the 'Sound Matching' portion of the reading subtest. Both items require the student to indicate whether two words, said aloud by the examiner, are the same or different. For each item, seven panelists noted that children from certain ethnic groups are unfamiliar with the differences between the critical sounds in the paired words.

Level B No biased items

Level C One item, from the 'Reading Comprehension, Passages' portion of the test. Seven panelists commented that the setting of the passage and the items of importance in the passage are not within the scope of experience of most minority children.

Level 1 Two items, one from the 'Reading Vocabulary' and one from the 'Reading Comprehension' portions of the test. The vocabulary items asked about a word that children of different ethnic groups would not experience often. The comprehension item contained a setting unfamiliar to many minority children.

Level 2 No biased items

Level 3 Three items, two from the 'Reading Vocabulary' portion and one from the 'Reading Comprehension' portion of the test. One vocabulary item dealt with a word considered outside the vocabulary of most black children, the other vocabulary item is the one mentioned earlier that could mislead students from Spanish-speaking backgrounds because of the Spanish root word. The comprehension item turned on a particular word judged to be an Anglo expression, and was based on a passage with a locale to which many minority children would not be exposed.

Level 4 No biased items.

After the biased items were identified, the data tapes provided by the publisher for the statistical de-biasing were rescored to determine the characteristics of the de-biased scores and how they related to the non-de-biased scores. The results of this final analysis are presented in Table 9-4.

Two notes of explanation appear necessary to conclude this description of the need, methods, and results of the de-biasing efforts of the CTBS. First, we must address the fact that only eight items (out of a population of 887 items) were ultimately identified as being biased. It may be concluded that this low incidence is due to the efforts of the publisher of the CTBS, first to develop tests that are minimally biased, and second to its own statistical de-biasing in order to weed out items that eluded the developers. It should be pointed out that at no time during the de-biasing process did the SDC team set compromising criteria for the identification of biased items, indeed, all the criteria were set prior to the study and appeared to be reasonable.

Table 9-4

Comparison of Non-De-Biased and De-Biased CTBS Reading Scores

Test Level	Non-De-Biased		De-Biased		r
Number and Ethnicity of Students	Mean	S.D.	Mean	S.D.	
Level A					
666 Kindergarten students	32 92	10 31	31 50	9 78	9984
222 Black	29 24	10 40	27 95	9 78	9982
222 Brown	31 36	9 27	30 03	8 80	9976
222 Other	38 17	9 05	36 51	8 63	9985
Level C					
693 Second-grade students	33 65	12 43	32 92	12 18	9996
231 Black	29 52	12.75	28 88	12 47	9995
231 Brown	32 81	11.71	32 06	11 49	9995
231 Other	38 62	11 08	37 83	10.81	9996
Level 1					
1 146 Third-grade students	41 60	19 19	40 73	18.72	9995
382 Black	34 66	16.58	33 91	16.20	.9993
382 Brown	36 83	16 49	36 08	16 10	9992
382 Other	53 31	18 75	52 21	18 18	9995
Level 3					
2 997 Seventh-grade students	42 19	18 40	40 59	17 68	9992
999 Black	34 39	15 52	33 20	14 92	9988
999 Brown	37 87	15 87	36 16	15 19	.9989
999 Other	54 52	17 05	52 42	16 46	9991

The second explanatory note is concerned with the relatively small mean differences in test scores resulting from the de-biasing effort and the high intercorrelation between the de-biased and non de-biased scores. Noting these (not unexpected) small changes, one might ask what was gained from the de-biasing effort. The statistics in Table 9-4 do not indicate how scores for different subgroups of students will be influenced, how growth indexes will be influenced, or how the significance and meaning of statistical comparisons will be influenced With the elimination of the biased items, however, we can be confident that however large the influences are, they will be in the direction of providing conclusions that will be fairer to minority and disadvantaged children.

THE USE OF ALTERNATE FORMS

Early in the planning of the schedule of tests, the Panel on Achievement Tests recommended that alternate forms of the tests not be employed in the SES. The panel was most concerned with the continuity of measurement over the years of the study. If longitudinal conclusions are to be drawn, then efforts must be made to ensure that similar aspects of the growth dimension are measured at each assessment point The utilization of alternate forms of the achievement test while addressing concerns that mere practice effects many account for findings also reduces the continuity of the measured outcomes. Empirical observations of alternate-form reliabilities of standardized achievement tests indicate that they tend to be about 10 lower than their corresponding internal-consistency or test-retest coefficients, Over a period of years with several retestings, the relatively small inconsistencies may influence the interpretation of the findings. Levine and Angoff (1958) present evidence supporting the conclusion that practice effects are generally smaller than 'inconsistency' effects in test-retest studies.

A more pragmatic reason for utilizing only one form of the achievement test involved the need to create a new vertical scale of growth The scaling procedures that are employed in the development of the vertical scale are highly dependent for their accuracy upon collinearity of the test levels being scaled Alternating test forms over the levels reduces the collinearity In addition, to the extent that the forms are not collinear, they will create different expansions or contractions within the scale. In this event, either different vertical scales would have to be constructed for the various sequences of forms, or the vertical scale would have to be 'smooth' over those differences. Based upon these considerations, we decided to employ only Form S of the CTBS for all test administrations.

ORDER AND LEARNING EFFECTS IN ADMINISTERING TWO LEVELS OF THE TEST

During the first year of the SES, each participating student was to be administered two levels of the standardized achievement test, a grade-appropriate (at-level) test and a test one level below grade-appropriate (below-level). This procedure is employed for two reasons. First, double testing should make possible the development of a highly accurate vertical scale of growth (see Report 10) Second, it will be possible to determine the best functional level of testing for each grade level at each school, so that testing during the second year and beyond can be at an optimal level for measurement. It is not unreasonable to expect, with a large representation of economically disadvantaged schools and students, that about half the schools participating in the study will employ below-level testing during the second year (see Hoepfner and Wang, this volume).

Where there is double testing, there is reason to believe that some phenomenon will influence scores on the second test that would not have influenced them had the second test been administered first. If, for example, second-test scores are depressed, we might expect fatigue or boredom to have played a role. If second-test scores are enhanced, perhaps learning or some sort of transfer has occurred. In more general terms, either outcome would indicate that the order in which a test is taken will influence the score earned. In the SES, the order of test administration should be determined to minimize the impact of the order effects upon the validity, acceptability, and logistics of the study.

From the validity standpoint, order effects had to be minimized for the at-level test, because those scores, converted to publisher-provided norms, were to be returned to the schools for their use. Order effects had also to be minimized so that they not be built into the vertical scale, thus giving an inaccurate picture of growth over levels of the test. From the point of view of acceptability, the sequence of test administered would have to be justified to school personnel, especially if it was the sequence that does not have an obvious rationale The logistics of the study are affected only by a decision to counterbalance—to employ different test administration orders at different classrooms, grade levels, or schools. Clearly, such differences complicate the packaging and record-keeping procedures in addition to the training of coordinators and classroom teachers. Actually, the complications caused by counterbalancing test order are sufficiently grave to render such a decision very undesirable from a logistical point of view.

The planning-year field test of the SES was, therefore, in part an effort to determine the test order(s) to be adopted for the first-year test administrations. Four schools each, within each of four categories, were selected for the field test. Northeast-urban-poor, Southeast-rural-poor, Southwest-urban-non-poor, and Northwest-rural-non-poor. At each school selected, two classrooms at grades 2, 4, and 6 were administered the achievement test; two randomly-assigned schools in each cell being tested with two reading tests, and two with two math tests Within each grade level of each school, one randomly assigned classroom was administered the below-level test first and the other was administered the at-level test first This balanced field test plan was designed to yield information on the effects of order upon the test scores. The tests were administered in a standardized way under the observation of a trained SES coordinator. The scores were converted to percentiles for ease of comparison, total score for reading and computation scale score for math. Table 9-5 presents the mean percentiles and the numbers of participating examinees under each testing condition for reading and for math.

Analyses of variance within each grade level for reading and for math revealed significant order effects in all cases. Statistical significance, however, must in this instance be interpreted in light of the meaningfulness (quantity) of effect observed. Table 9-5 was supplied for this purpose.

Table 9-5

Analysis of Order Effects on Reading and Math Percentile Scores

Grade	Order					
	N	Below-Level First	At-Level Second	N	At-Level First	Below-Level Second
Reading						
2	178	50.44	59.16	174	54.66	48.37
4	185	42 42	44.67	181	45 81	44.75
6	196	49.43	47.51	193	49.05	46.93
All	559	47.43	50.[illegible]8	548	49.76	46.67
Math						
2	189	37 65	42.01	206	45.92	50.82
4	215	49 95	50 08	218	60.64	57.48
6	214	55 28	50.92	213	51.88	55.54
All	618	48.03	47.90	637	52.00	54.68

Comparing mean reading percentiles for the same level over the different orders, one can detect small and inconsistent differences. The largest difference is 4.5 percentiles for the at-level test, higher when administered second. Over all grades, it can be seen that the below-level test mean is reduced by .76 percentile when it is administered second, while the at-level test mean is increased by .52 percentile. The average difference in percentile means between the two levels is 4.19 percentiles in the below-then-at order, and 3.09 percentiles in the at-then-below order. Considering the rather large standard error for percentile conversions, these differences may all be meaningless. From these summary findings it would appear that the reading test could safely be administered in any order.

Comparing mean math percentiles for the same level over different orders, the differences seem consistently larger, although just as inconsistent. The largest difference in Table 9-5 is 13.17 percentiles for the below-level test, higher when administered second. Over all grades it can be seen that the below-level test mean is increased by 6.65 percentiles when it is administered second, while at the at-level test mean is reduced by 4.10 percentiles. The average difference in percentile means between the two levels is 2.88 percentiles in the below-then-at order, and 3.89 percentiles in the at-then-below order.

Because these data do not provide an unequivocal answer to the question regarding order of test administration, it was decided to consider each order separately to see what advantages and disadvantages would result. Weighing the advantages and disadvantages in each test order, it was decided to test in a constant below-then-at order. This order would meet with the greatest school personnel approval, would reduce test trauma in first-graders, and would result in very few and very small technical advantages.

LEVEL OF SCORES TO BE ANALYZED

Administration of the achievement tests can result in several different types of scores. In the case of the CTBS, several varieties of types of scores can be considered. As examples, one might wish to study total achievement scores, skill-area achievement scores, sub-scale scores, or even item scores. We elected to employ skill-area scores (reading total and math total) rather than more general or more constituent scores. This decision was made on the basis of three considerations. acceptability, vertical-scale construction, and collinearity among dependent variables.

Score Acceptability. On the basis of conversations with many educators, it was concluded that few if any teachers or school principals would be concerned that the study employed skill-area scores rather than total or sub-scale scores. Although instruction is presumably targeted at more specific levels, its effects spread across specific learning objectives. Reading teachers, and most materials used to teach reading, stress both vocabulary building and comprehension improvement—most frequently together, in math the stress is on concepts largely as an aid to computation This argument confirmed us in our decision not to observe the traditional approach, but to analyze achievement scores at the higher and more easily understood conceptual level.

Vertical Scale Construction. The development of adequate vertical scales requires that the achievement-test data collected meet some basic assumptions:

That the scores have a sufficiently fine metric such that many gradations are possible: the more gradations, the more accurate the intra-level matching can be.

That the scores over the entire vertical range are linearly related, i.e., the scores measure the same thing or the same collection of things.

That the scores are reliable measures of what they measure, so that scale extrapolations from scores can be made with confidence that the score belongs where it is placed on the vertical scale.

We studied the publisher's data to see if the use of subtest scores could be justified and, if so, what problems might arise that would be dangerous to the conclusions of the study. Tables 9-6 and 9-7 provide the scores and the number of items for each score that would be used under skill-area total and under sub-scale score conditions.

The scaling of skill-area total scores will be fairly straightforward. The number of items (gradations of metric) is large, and the even columns indicate that conceptually (at least) there is commonness of content (linearity).

Table 9-7 presents quite a different picture. The number of items of some of the sub-scale scores is very small, setting severe limitations upon the accuracy of any vertical scale that is built upon them. Further, the uneven columns indicate that there are problems as to which scores would be scaled to which other scores at different levels. For example, while there appear to be no problems at and above Level 1, at the early levels we would have to

Table 9-6

Vertical Scales if Skill-Area Totals Are Used

Test Level	Reading Totals	Math Totals
A	Reading total: 49 items	Math total: 26 items
B	Reading total: 43 items	Math total: 46 items
C	Reading total: 51 items	Math total: 40 items
1	Reading total: 85 items	Math total: 73 items
2	Reading total: 85 items	Math total: 73 items
3	Reading total: 85 items	Math total: 73 items
4	Reading total: 85 items	Math total: 73 items

Table 9-7

Vertical Scales if Sub-Scale Scores Are Used

Test Level	Sub-Scales	
	Reading	
A	Sound Matching: 28 items	Letter Sounds: 21 items
B	Word Recognition II: 19 items	Reading Comprehension: 24 items
C	Vocabulary: 33 items	Comprehension Passages: 18 items
1	Vocabulary: 40 items	Comprehension: 45 items
2	Vocabulary: 40 items	Comprehension: 45 items
3	Vocabulary: 40 items	Comprehension: 45 items
4	Vocabulary: 40 items	Comprehension: 45 items
	Math	
A	Mathematics: 26 items	
B	Concepts: 14 items	Computation: 32 items
C	Concepts: 12 items	Computation: 28 items
1	Concepts: 25 items	Computation: 48 items
2	Concepts: 25 items	Computation: 48 items
3	Concepts: 25 items	Computation: 48 items
4	Concepts: 25 items	Computation: 48 items

choose, according to some criterion (unknown), which of the reading sub-scales belongs to the 'vocabulary' vertical scale and which to the 'comprehension' vertical scale. The problem is even worse when one observes that in the math area there is only one short test at Level A: to which vertical scale, 'concepts' or 'computation,' does it get assigned? (It shouldn't be assigned to both, because findings will be totally collinear then, and if the scale is split, there will be so few items that the vertical scale will be exceedingly unreliably constructed.)

Of course, the continuity of the vertical scales cannot be judged solely on the basis of the names attached to the scales For that reason Table 9-8 presents the empirical intercorrelations between scales in Levels 1 through 4, using both skill-area scores and sub-scale scores.

Notice that in Table 9-8, even disregarding the problems of linearity between levels at the lower grades, the inter-correlations between levels using skill-area scores are consistently higher generally by an average of about +.045. Keep in mind that this small difference does not include any differences at the critical lower levels. (The publisher did not provide intralevel inter-correlations because of the difficulty of matching; use of the sub-scale scores would compound this difficulty)

Another way of evaluating the effectiveness of skill-area vs. sub-scale total scores would be to look at their internal reliabilities. These reliability estimates are provided in Table 9-9.

The average internal reliability for the skill-area totals is .9229; while that for the sub-scale totals is 8959, for an average difference of +.0270 in favor of the skill-area total scores.

A closely related approach to the relative evaluation of the score types involves inspection of the standard errors of measurement (SEMs) of the scores. The SEMs reported in Table

Table 9-8

Intra-Level Intercorrelations Among Scales

Test Levels	Skill-Area Total Scores			
	Reading		Math	
1 and 2	86		75	
2 and 3	87		78	
3 and 4	84		.83	
	Sub-Scale Total Scores			
	Vocabulary	Comprehension	Concepts	Computation
1 and 2	79	80	72	.75
2 and 3	76	81	77	78
3 and 4	79	76	76	.83

Table 9-9

Internal Reliabilities of Scales

Test Level	Skill-Area Total Scores: Reading	Skill-Area Total Scores: Math
A	.91	.81
B	.91	90
C	93	.89
1	.96	.95
2	.96	.92
3	96	.94
4	.94	.94

Test Level	Sub-Scale Total Scores: Reading	Sub-Scale Total Scores: Math
A	Sound Meaning: .91 Letter Sounds. 88	Mathematics .81
B	Word Recognition II. 87 Reading Comprehension: .91	Concepts: .83 Computation: .90
C	Vocabulary: .93 Comprehension. .92	Concepts: .85 Computation: .89
1	Vocabulary: .92 Comprehension: .94	Concepts: .89 Computation: .95
2	Vocabulary: .92 Comprehension: .92	Concepts: .82 Computation: .92
3	Vocabulary: .93 Comprehension: .92	Concepts: .85 Computation: .94
4	Vocabulary: .90 Comprehension: 90	Concepts: .84 Computation: .93

9-10 are for vertical-scale scores, as they presumably have more nearly equal units of measurement.

Summarizing Table 9-10 in terms of overall averages would be misleading, but it is important to notice that the SEMs in the rightmost column are considerably smaller than those in the left columns in eight of the twelve comparisons. These SEMs, of course, will underly all the analyses and interpretations of achievement outcomes in the SES.

To summarize the concerns regarding the development of adequate vertical scales, the skill-area scores exhibit finer gradation, greater reliability, and greater between-level correlations than do the sub-scale scores. In addition, it is possible to assign scores into the

Table 9-10

Standard Errors of Measurement of Skill-Area and Sub-Scale Total (Scaled Scores)

Test Level	Sub-Scales				Skill Area
			Reading		
A	Sound Matching:	3 01	Letter Sound:	3.59	—
B	Word Recognition II:	7.72	Reading Comprehension:	6.01	8.33
C	Vocabulary:	15 90	Comprehension Passages:	9.67	9.93
1	Vocabulary:	19.77	Comprehension:	20 16	14.04
2	Vocabulary:	22 83	Comprehension:	26 30	17:52
3	Vocabulary:	27.99	Comprehension:	31.14	20 27
4	Vocabulary	34 80	Comprehension.	37.75	24.81
			Math		
A	Mathematics	13.45			
B	Concepts:	11.26.	Computation:	12.47	12.77
C	Concepts:	11.37	Computation	14.59	13.14
1	Concepts.	23 67	Computation:	10:95	9.80
2	Concepts.	34.81	Computation:	17.84	14.69
3	Concepts:	37 68	Computation:	20.59	16.90
4	Concepts:	44 28	Computation:	25.94	21.20

vertical scales very unambiguously in the case of the skill areas. At the lower test levels there is no obvious route to assign sub-scale scores to vertical scales.

Collinearity of the Dependent Variables. It is probably true that all dependent variables in the SES will be collinear to some degree. It was our desire, however, to select measures that would be minimally collinear. The reason for this, of course, is that with non-collinear dependent variables, all findings can be directly interpreted; while with collinear ones, discrepancies between findings cause problems. The best indication of the collinearity to be expected is to examine the intercorrelations between the dependent variables when sub-scale total scores are used. These intercorrelations, both raw and corrected for unreliability, are reported in Table 9-11.

The average intercorrelation, corrected for unreliability of the component scales, is .81. This means that 66 percent of the reliable variance of any of the dependent variables would be shared by its parallel dependent variable, on the average. The problem with such collinearity is not that the analyses are all close to duplicates of each other, but that when differences are observed between two paired dependent variables, the chances are that the differences are artifactual and should not be interpreted. In this event, of course, the contention that the study will provide information that can be interpreted differently for each subject-matter component causes it to lose considerable credibility.

Table 9-11

Raw and Corrected Intercorrelations Between Reading Sub-Scale and Math Sub-Scale Total Scores as Indexes of Collinearity

Test Level	Reading Intercorrelations		Math Intercorrelations	
	Raw	Corrected	Raw	Corrected
A	.51	57	—	—
B	74	83	70	81
C	62	.67	59	.68
1	83	.89	73	.79
2	.82	89	69	.79
3	84	91	79	.88
4	82	91	.80	91

The argument for the use of total scores rather than sub-scale scores is strengthened by providing empirical information on the collinearity not only of the sub-scale scores but also of the 'gain' scores derived from the sub-scales. Table 9-12 provides critical information based on second-year data from the ESAA study (Ozenne, 1976). The correlations between the various sub-scale gain scores are considerably lower than those between the sub-scale scores. This should not be considered as arguing against the collinearity of the gain scores because the reliability of those gains is very questionable. (Theoretically, the gains are equal to the posttest scores minus all the true variance that they shared with the pretest scores, meaning that their true variance has been reduced much more than their error

Table 9-12

Intercorrelations (Collinearity) Between Sub-Scale Raw Scores* and Raw Sub-Scale Gain Scores**

Grade	Reading		Math	
	Sub-Scale	Sub-Scale Gain	Sub-Scale	Sub-Scale Gain
4	814	351	767	644
5	803	551	746	555
6	814	586	785	.517

*Scores based on appropriate second-year posttest levels of the California Achievement Tests, Reading Vocabulary and Comprehension, and Mathematics Concepts and Computation. Intercorrelations are corrected for unreliability of subtests.

**Gain scores are computed as residuals, with pretest predicting posttest of alternate-form sub-scale scores Intercorrelations are corrected for unreliability caused by pre-post time span, alternate forms, and subscale unreliability, but not for reduction in true-score variance

variance This problem has been called a paradox in the literature, and is here to haunt us now.)

Quite to the contrary, considering the general unreliability of residual gain scores, the coefficients of collinearity reported above appear to be very large and are interpreted as arguing again for utilization of skill-area total scores.

THE TESTING PLAN

On the basis of the information accumulated through the test screening procedure, the visits to the test publishers, and other supporting analyses described above, the following test plan was developed.

Fall of the First Year. Each student was to be tested with the grade-appropriate (at-level) test and with the test one level below appropriate (below-level). Just which level was at-level or below-level was largely to be determined by the grade-level recommendations of the publisher, but in some cases additional information had to be considered. On the assumption that each student would be administered two levels of the test, it was important to assure that the lower level of the test would not be severely 'topped out' and that the higher level of the test would not be severely 'bottomed out.' Table 9-13 presents average item difficulties for all the levels and for grades Kindergarten through 6. Average item difficulties were extrapolated in a linear manner for those grades and levels for which empirical data were not available. Inspection of the item difficulties, and secondary inspection of the percentile equivalent of 'chance' scores, resulted in decisions regarding which level would be considered at-level for each grade and which would be considered below-level. The tests in the heavy lines of Table 9-13 are the tests that were selected.

Based on the study of order effects, the testing plan shown in Table 9-14 was proposed for the fall of the first year. The fall testing plan provided for the optimum ordering of test administrations so that the tests would not interact, and also provided days of non-testing, which the field-test observers considered critical for obtaining valid test results.

Spring of the First Year. The testing plan proposed in Table 9-14 for the spring of the first year closely parallels that of the fall, with the exception that the levels of some of the standardized achievement tests change to meet the growing skills of the students.

After the fall testing of the first year, the below-level and at-level test scores for each grade level at each school were examined. On the basis of examination of means and standard deviations, we could determine the appropriate levels for later years.

Later Years. The testing plan for the study years after the first year is also presented in Table 9-14. Notice that because there was no longer a need for double administration of the standardized achievement test that the number of testing days has been reduced.

196

Table 9-13

Scale and Item Characteristics* (Empirical and Linearly Extrapolated) of the Levels of the CTBS at Various Testing Grades and Times**

Level and Grade Range	Scales***	No. of Items Alternatives	Mins.	Kindergarten Fall	Kindergarten Spring	Grade 1 Fall	Grade 1 Spring	Grade 2 Fall	Grade 2 Spring	Grade 3 Fall	Grade 3 Spring	Grade 4 Fall	Grade 4 Spring	Grade 5 Fall	Grade 5 Spring	Grade 6 Fall	Grade 6 Spring
Level A K 0-1 3	Reading	49 03	39	57 28	74 05	91 05	99										
	Math	26 03	28	50 19	65 04	80 04	95										
Level B K 6-1 9	Reading	43 03	39		34 47	38 41	68 07	98									
	Math	44 03	45		36 47	40 42	60 12	99									
Level C 1 6-2 9	Reading	51 04	36			32	46 17	59 07	73 02	87	99						
	Math	41 04	47			29	42 18	54 11	67 03	80	93						
Level 1 2 5-4 9	Reading	85 04	46					33 39	41 24	49 16	57 08	62 06	68 04	74	80	86	
	Math	73 04	55					30 44	42 25	55 13	67 06	73 04	79 03	85	91	97	
Level 2 2 5-4 9	Reading	85 04	46								39	45 14	50 09	55 07	61 05	64 04	68 04
	Math	73 04	55								40	46 11	52 07	53 05	64 04	66 03	69 03
Level 3 2 5-4 9	Reading	85 04	46													50 12	53 11
	Math	73 04	55													46 16	49 13

*First entry is average item difficulty index, second is percentile equivalent of a 'chance' score

**Levels enclosed in lines are those suggested for the SES, based on considerations of samples, item difficulties, and scaling requirements

***Sub-scales are listed in Table 9-10

Table 9-14

Testing Plan for the SES

Day	Grade 1	Grade 2	Grade 3	Grade 4	Grade 5	Grade 6
				Fall of the First Year		
Week 1						
Monday Tuesday	PT	PT/SAM	PT/SAM	SAM	SAM	SAM
Wednesday Thursday	CTBS-R-A	CTBS-R-B	CTBS-R-C	CTBS-R-1	CTBS-R-1	CTBS-R-2
Friday	CTBS-M-A	CTBS-M-B	CTBS-M-C	CTBS-M-1	CTBS-M-1	CTBS-R-2
Week 2						
Monday Tuesday				PAS	PAS	PAS
Wednesday Thursday	CTBS-R-B	CTBS-R-C	CTBS-R-1	CTBS-R-2	CTBS-R-2	CTBS-R-3
Friday	CTBS-M-B	CTBS-M-C	CTBS-M-1	CTBS-M-2	CTBS-M-2	CTBS-M-3
				Spring of the First Year		
Week 1						
Monday Tuesday	PT/SAM	PT/SAM	PT/SAM	SAM SAM	SAM SAM	SAM SAM
Wednesday Thursday	CTBS-R-B	CTBS-R-C	CTBS-R-C	CTBS-R-1	CTBS-R-1	CTBS-R-2
Friday	CTBS-M-B	CTBS-M-C	CTBS-M-C	CTBS-M-1	CTBS-M-1	CTBS-M-2
Week 2						
Monday Tuesday				PAS	PAS	PAS
Wednesday Thursday	CTBS-R-C	CTBS-R-1	CTBS-R-1	CTBS-R-2	CTBS-R-2	CTBS-R-3
Friday	CTBS-M-C	CTBS-M-1	CTBS-M-1	CTBS-M-2	CTBS-M-2	CTBS-M-3
				After the First Year		
Week 1						
Monday Tuesday	PT/(SAM)	PT/SAM	PT/SAM	SAM	SAM	SAM
Wednesday Thursday	CTBS-R-X	CTBS-R-X	CTBS-R-X	CTBS-R-X	CTBS-R-X	CTBS-R-X
Friday	CTBS-M-X	CTBS-M-X	CTBS-M-X	CTBS-M-X	CTBS-M-X	CTBS-M-X
Week 2						
Monday				PAS	PAS	PAS

Note: PT stands for 'Practice Test'; SAM for 'Student Affective Measure', CTBS for 'Comprehensive Tests of Basic Skills,' followed by R for 'Reading' or M for 'Mathematics' and by level designation, PAS for 'Practical Achievement Scale.' Parentheses around SAM indicate that it is not administered during the fall testing, but is during the spring. The X indicates that the CTBS level is not known, but will be determined as a functional level for each grade at each school.

CHAPTER 10. THE FEASIBILITY OF USING CRITERION-REFERENCED TESTS IN THE SUSTAINING EFFECTS STUDY

Jacqueline Kosecoff and Arlene Fink*

The measurement needs of the Sustaining Effects Study (SES) and the theoretical underpinnings and practical limitations of criterion-referenced testing systems were exhaustively compared in order to determine the feasibility of employing such an approach in a nationwide study. In general, the practical limitations of criterion-referenced tests and the level of effort needed to remedy those limitations led to the conclusion that the use of such instrumentation is not feasible.

Criterion-referenced tests (CRTs) have become increasingly popular among educators and psychometricians. Perhaps the most important reason for their appearance and widespread acceptance can be traced to the new ways that had to be found to measure the effects of the educational reforms of the 1950's and 1960's. During those decades, the conventional school curriculum was declared in need of reform, and a reassessment of the goals and objectives of American education was made (Hofstadter, 1963, Davis and Diamond, 1974, Cronbach and Suppes, 1969). Innovative courses of study and instructional technologies were subsequently developed, and programmed learning and individualized instruction became commonly used teaching techniques. New ways of assessing student performance were needed that would correspond to the teaching innovations.

Educators have traditionally relied on paper-and-pencil achievement tests to measure learning, so it was natural for them to turn to test theoreticians to provide them with alternative ways of interpreting performance on measures of education achievement for the new curricula and methods of instruction. The psychometricians responded by pointing to two basic ways of assigning meaning to test scores. The first involved comparing the performance or behavior of one person or group with another person or group, and the second involved describing what a person or group can do or can be expected to do. Glaser (1963) referred to these two ways of giving meaning to test scores as norm-referenced and criterion-referenced, and recommended criterion-referenced score interpretations for the reformed curriculum and instruction.

The reaction to CRTs was enthusiastic from the start. Because they provide score interpretations in terms of the achievement of specific and measurable skills and behaviors, CRTs have had appeal to those directly responsible for the education of students and the development and evaluation of educational programs. They also have had appeal to teachers who found the results of standardized tests inadequate to assist them in planning lessons, and to many educators and psychologists who judged standardized, norm-referenced tests to be unfair and even biased against individuals from under-privileged and

*Drs Kosecoff and Fink prepared this report under subcontract to SDC minority groups. Finally, because the criterion-referenced approach was new, people saw it as an opportunity to improve on some of the mistakes they perceived to be built into norm-referenced testing

The popularity of CRTs and their sanction by both theoreticians and practitioners has led to their frequent use for instructional diagnosis and placement, and for measuring student achievement on educational tasks or objectives. In addition, CRTs are now being suggested or used for other purposes, such as the evaluation of educational programs and the National Assessment of Educational Progress (Wilson, 1974). In fact, many state and federal agencies have specifically required evaluators to justify their selection of standardized rather than CRT measures.

To determine if CRTs are appropriate for the SES, we first examined the theory that structures their development and validation to determine whether, on theoretical grounds alone, CRTs are suitable or not suitable for national effectiveness evaluations. The next step was to identify a set of criteria for selecting tests that are appropriate for such an evaluation; included within the set of criteria was the stipulation that the test be able to provide scores amenable to CRT interpretation. We then reviewed currently available CRTs, using the criteria. Finally, based on the theoretical examination and the review, we formed conclusions concerning the use of CRTs in the SES. This investigation is organized into four parts: the effectiveness evaluation context and the SES, a theoretical examination of criterion-referenced testing; a review of currently available CRTs; and conclusions.

THE EFFECTIVENESS EVALUATION CONTEXT AND THE SES

The evaluation of an educational program involves the use of specific procedures that result in an appraisal of the program's basic merit, and provides information about the nature and quality of the program's goals, outcomes, impact, and costs (Fink and Kosecoff, 1980).

Evaluation Contexts. There are two contexts in which evaluations of educational programs are conducted. In one context, an evaluation is conducted to *improve* a program, and the evaluation's clients are typically the program's organizers and staff. In the second context, an evaluation is conducted to measure the *effectiveness* of a program, and the evaluation's clients are typically the program's sponsors. The context for an evaluation is determined by the information needs of the individuals and agencies that must use the evaluation information.

An evaluation is performed in an improvement context when the evaluation's clients are concerned with finding out precisely where a change would make the program better. Typically, the organizers of a still-developing program require this kind of information so that they can modify and improve the program. On the other hand, an evaluation is conducted in an effectiveness context when the evaluation's clients are particularly concerned with determining the consistency and efficiency with which the program achieves desired results. Those individuals who sponsor a program's development, or who are interested in using the program, require this kind of information about a well established program's outcomes and impact.

In an effectiveness evaluation, the evaluator usually assumes a more global and independent stance toward the program than in an improvement context. In addition, the effectiveness evaluator usually makes use of powerful, experimental design strategies that permit comparisons, rely on empirically-validated and standardized instruments, and employ statistical and other analytic methods that allow inferences regarding the program's comparative value.

Effectiveness evaluations of educational programs are conducted to appraise a program's overall impact and worth, and, if it is deserving, to certify it as being able to produce certain outcomes efficiently. This type of evaluation is frequently designed for programs that are relatively well-developed and stable, having defined purposes and fixed profiles. Other names for evaluations of this sort include summative evaluation (Scriven, 1967) and outcome evaluation (Klein et al., 1971).

The major focus of the SES is the Title I program, originally funded through the Elementary and Secondary Education Act of 1965. Title I represents one of the largest attempts by the federal government to assure equal educational opportunity for the disadvantaged. Because the purpose of the evaluation is to assess this established and long-lived program's impact in terms of costs and learning, it fits within the framework of an effectiveness evaluation It is generally agreed (e.g., Alkin et al., 1974) that information-collection strategies for projects like the SES should rely upon instruments that are known to provide relevant information and that have been demonstrated to be valid and reliable for the target population.

A THEORETICAL EXAMINATION OF CRITERION-REFERENCED TESTING

This section presents an investigation of the theoretical issues involved in the development and validation of CRTs. These issues include: defining CRTs; formulating and generating CRT objectives, items, and score interpretation schemes; establishing item and test quality; and examining the use of classical indexes of reliability and validity.

Definition: A criterion-referenced test is one that is designed to provide a measure of the extent to which educational purposes or tasks have been achieved. All CRTs share several features in common:

- They are based on clearly defined educational tasks and purposes.
- Test items are specifically designed to measure the purposes and tasks.
- Scores are interpreted in terms of the attainment of preset criteria or levels of competence with respect to the purposes and tasks.

All definitions of CRTs involve reporting test scores in terms of achievement of educational tasks.

The difference between criterion-referenced tests and norm-referenced tests lies mainly in the metric used to describe their scores. Criterion-referenced tests report scores, using metrics such as mastery or percentage of an objective achieved, that are intended to measure levels of competence or achievement in terms of a performance criterion. Norm-referenced tests report scores, using metrics such as percentiles and stanines, that are intended to permit comparisons or rankings. All other differences between norm-referenced and criterion-referenced tests, such as the way each is developed and validated, are derived from the need to produce tests that permit the appropriate score interpretation.

Development of Criterion-Referenced Tests

Formulating and Generating Objectives. A clearly defined set of educational tasks and purposes (objectives) is required for CRT development. Such objectives are selected in at least six ways:

1. *Expert judgment*. On the basis of knowledge and experience in the field, experts assess which educational tasks and purposes are the most important to measure.

2. *Consensus judgment*. Various groups such as community representatives, curriculum experts, teachers, and/or school administrators decide which educational tasks and purposes they consider to be the most important to measure (Klein, 1972; Wilson, 1973)

3. *Curriculum analysis*. A team of curriculum experts analyzes a set of curriculum materials in order to identify, and, where necessary, to infer the educational tasks and purposes that should be the focus of the test (Baker, 1972)

4. *Expert analysis of subjects*. An in-depth analysis is made of a subject area to identify all knowledge and skills that must be acquired if the area is to be learned well (Glaser and Nitko, 1971; Nitko, 1973).

5. *Theories of learning and instruction*: The literature is reviewed and/or experts are consulted to formulate series of hierarchies of educational tasks and purposes based upon the results of psychological theory and research (Keesling, 1975).

6. *Empirical studies*. Experiments are conducted to identify objectives that are important because the skills and knowledge represented are inherently essential.

No matter how they are derived, educational tasks and purposes are usually called objectives or behavioral objectives. However, it should be noted that these terms have a precise meaning to educators:

> An objective is an *intent* communicated by a statement of what the learner is to be like when he has successfully completed a learning experience (Mager, 1962).

Developers of CRTs do not always use this definition in its purest sense. Rather, they often use *objective* to refer to the content that is supposed to have been learned (e.g., equivalent and nonequivalent sets in sixth-grade math) and sometimes also include the behaviors a student is supposed to exhibit (e.g., naming the first five Presidents of the United States).

Another issue concerning educational tasks and purposes, e.g., objectives, relates to the rules needed for writing objectives and how broadly or narrowly they should be stated. Formal rules for generating and stating objectives are needed to ensure the uniformity, manageability, and comprehensiveness of the set of objectives or domain that the CRT measures.*

Still another issue deals with how a domain is organized. The objectives for a single domain can be grouped by grade levels; they can be organized according to major content areas; and/or they can be arranged into a hierarchy according to the complexity of the behaviors involved or the order of instruction.

Formulating and Generating Items. Once the objectives for the CRT have been chosen, the next step is to construct and/or select test items to measure them. This is one of the most difficult steps in the developmental process because of the vast number of test items that might be constructed for any given objective, even when objectives are relatively narrowly defined (Klein and Kosecoff, 1973). For example, consider the following objective: "The student can compute the correct product of two single-digit numerals greater than zero where the maximum value of this product does not exceed 30." The specificity of this objective is quite deceptive since there are 55 pairs of numerals that meet this requirement and at least ten different item types that might be used to assess student performance.

Further, each of the resulting 550 combinations of pairs and item types could be modified in a variety of ways that might influence whether they have been answered correctly. Some of these modifications are:

1. Vary the sequence of numerals (e.g., 5 then 3 versus 3 then 5).

2. Use different item formats (e.g., multiple-choice versus completion).

3. Change the mode of presentation (e.g., written versus oral).

4. Change the mode of response (e.g., written versus oral).

It is evident that a highly specific objective could have a potential item pool of well over several thousand items (Hively, 1970; Hively et al., 1973; Bormuth, 1970). Several factors influence the number of items constructed for each objective, e.g., the amount of testing time available and the cost of making an interpretation error (such as saying that a student has achieved mastery when he or she has not). For some objectives, many items are need-

*The set of objectives that a CRT measures is sometimes called a 'domain' or 'universe' of content (Skager, 1975, Cronbach, 1971). However, the term domain is used by others to mean the rules for generating test items to measure a specific objective (Hively, et al., 1973).

ed to obtain a stable estimate of a learner's performance, whereas for other objectives fewer items will suffice.

A related issue in the construction and generation of CRT items is the degree to which the items should be sampled with respect to their difficulty and content coverage within an objective. It is a well known and frequently used principle of test construction that even slight changes in an item can affect its difficulty. The extent to which the items within an objective are sampled with respect to difficulty has a direct bearing on the interpretation of the scores obtained. In other words, if only the most difficult test items are used, the phrase 'achievement of the objective' has a very different meaning than if the test items are samples taken from the full range of difficulties.

Another issue, the instructional dependence of a CRT, refers to the extent to which the CRT is designed for use with a specific educational program (Baker, 1972, Skager, 1973). CRTs with a greater degree of instructional dependence have objectives and test items that are associated with a particular curriculum or set of educational materials and tecnhiques. CRTs with a smaller degree of instructional dependence, on the other hand, contain objectives and test items that are not necessarily associated with the specific skills or content of an educational program. However, such CRTs still may have been developed from several educational programs and, consequently, have objectives and items that reflect the emphasis of these programs. Conversely, CRTs with no instructional dependence are based on a domain of content and behaviors that are independent of any educational program, and, therefore, can be used to compare several different educational programs. (The development of this latter kind of CRT is, in fact, similar to that employed in the better standardized norm-referenced tests.)

Consideration of the various issues involved in item generation for CRTs has produced different strategies for generating and constructing items, namely:

- *Panel of experts.* A group of measurement and curriculum experts decided which items to use based on their knowledge of an experience in the field (Zweig, 1973).
- *Content/process matrix.* Basically a variation of the classical test construction technique, this approach involves developing a matrix of contents and behaviors (or tasks) to be assessed. Items are then systematically sampled within this matrix and perhaps along a third continuum of item difficulty as well (Wilson, 1973).
- *Systematic item generation:* Basic item forms or specifications are developed for each objective that define the range of item difficulties, all the relevant contents and behaviors, and stimulus and response characteristics of items that can be used to assess the objective (Hively, 1970, Hively et al., 1973, Cronbach, 1971, Skager, 1973; Popham, 1975).

Formulating Score Interpretation Schemes. The uniquely distinctive feature of a CRT is its ability to provide a means for describing what an individual or group can do, know, or feel without having to consider the skill, knowledge, or attitude of others. Consequently, CRT scores are reported and interpreted in terms of the level of performance obtained with respect to the objective(s) or domain on which the CRT is based. This type of score reporting is very different from that used for norm-referenced tests in which scores are reported in terms of the performance of other individuals or groups.

It should be noted that scores on CRT tests need not be limited to just a CRT interpretation. Other interpretations can also be provided to expand upon the CRT interpretation (Klein, 1970; Cronbach, 1970, Ebel, 1972). An example of one way of combining criterion- and norm-referenced information is. This school had an average score of 5 out of 10 on the objective (a CRT interpretation), which is one standard deviation below the national average of 7 out of 10 (a norm-referenced interpretation). The idea of using both types of interpretations is not new and does not reduce the theoretical soundness of the score interpretation Combining score interpretations is particularly useful for describing what a student can be expected to be able to do and how exceptional or typical the performance of the student is.

Some of the different scores that can be interpreted in a CRT sense are:

- *Actual score* The number or percentage of items correct on a given objective
- *True score* An individual's or group's true level of performance on an objective, referring to the portion of the total universe of items for an objective that an individual or group could answer correctly. The true score is the number of items that an individual or group would pass if every possible item were tested.
- *Mastery of an objective*. The achievement of a preset criterion level of performance. To be legitimate, the criterion level should be meaningful and preferably empirically justifiable. For example, a criterion level of 7 out of 10 items has meaning if systematic study has shown that those who reach this level can actually do something that others who have not reached it cannot do, or if baseline data show that the average student achieves this level.
- *Performance time*. This is the time it takes, in class hours or calendar days, for a student to achieve a given performance level.
- *Level readiness*. This score reflects the probability that the student is ready to begin the next level of instruction (this may be based on both the number of items correct and the pattern of answers given to these items).
- *Item difficulty*. The difficulty of an item is measured by the percentage of students who pass the item. (This score is given most often when only one item is tested per objective; for example, National Assessment of Educational Progress.)
- *Total objectives mastered*. The number of objectives passed or mastered by an individual or group.
- *Total individuals who passed*. The number of individuals or groups who passed or mastered each objective.

Validation of Criterion-Referenced Tests. When construction of the objectives and test items is complete, the CRT must be analyzed and validated. This process can involve giving the test to students and studying their responses (response data), or relying on a view of the test by experts (judgmental data). There is much ambiguity about the procedures appropriate for analyzing and validating CRTs. There are, nevertheless, several measures (or dimensions) used for item and quality testing that are relevant to CRT validation, and that have associated with them review procedures, data-collection strategies, experimental designs, and statistical indexes.

Establishing Item Quality. Following are several commonly considered dimensions of item quality:

- *Item-objective congruence.* A test item is considered good if it measures or is congruent with the objective it is supposed to assess. Item-objective congruence can be established by using judgmental data.

- *Equivalence (internal consistency within objectives).* A test item is considered good if it behaves like other items measuring the same objective. The concept is similar to item objective congruence, but its proper use depends on response data. Equivalence is usually measured by computing the biserial correlation between the score on an item and the total score on all items measuring that objective.

- *Stability (over time).* An item is considered good if examinee performance is consistent from one test period to the next in the absence of any special intervention (e.g., instruction is an intervention that can change performance). Stability involves response data and can be measured by using a *phi* coefficient that correlates scores on the item from two different occasions.

- *Sensitivity to instruction.* An item is considered good if it is sensitive to instruction, that is, if there is a discrimination in responses to the item between those who have and those who have not benefited from instruction. This measure of item quality is usually computed for CRTs that are linked to particular educational programs and it requires response data.

- *Cultural/sex bias.* An item is considered good if it leads to accurate inferences about the knowledge, skills, or other attributes of an individual or group. Bias can be assessed using either judgmental or response data.

Establishing Test Quality. There are seven dimensions commonly used to express the quality of a CRT:

- *Test-objective congruence.* Similar to item-objective congruence, test-objective congruence assesses the extent to which the total test or subtest measures the relevant objective. Test-objective congruence is usually determined by using judgmental data.

- *Equivalence (internal consistency).* Test equivalence measures the homogeneity of test items for an objective, that is, how coherently the test items assess the particular objective. This can be measured by using split-half correlation, Kuder-Richardson formulas, or coefficient *alpha*.

- *General stability (test-retest, or alternate forms):* A test is stable to the extent that examinee responses are consistent from one test period to another or across alternate forms of a test in the absence of any intervention.

- *Quantitative stability (number of items per objective and number of objectives per domain):* There are two levels at which this type of stability for a CRT can be estimated. At the first level, a determination is made of the number of items that should be tested in order to obtain a stable score on an objective. At the second level, a determination is made of the number of objectives that should be tested in order to obtain a stable estimate of performance on the domain. Stability can be estimated with response data using correlation techniques and/or Bayesian models (Novick and Lewis, 1974).

- *Sensitivity to instruction.* Sensitivity to instruction refers to a test's ability to discriminate between those who have and those who have not benefited from instruction. This type of measure of test quality is usually obtained for CRTs that are linked to a specific educational program. It can be measured using response data by comparing scores of those who have and those who have not received instruction.

- *Cultural/sex bias.* Bias in measurement occurs when characteristics of the test, the testing process, or the interpretation of test results lead to inaccurate inferences about the knowledge, skills, or other attributes of individuals or groups (Anderson et al., 1975). This can be measured by ANOVA or regression techniques using response data, or by expert review using judgmental data.

- *Criterion validity.* Criterion validity establishes the meaningfulness of the criterion in terms of which CRT scores are interpreted. Establishing criterion validity is either a one-step or a two-step process. The first step involves assessing the meaningfulness of the domain. That objectives have been selected and organized to be, in themselves, educationally significant, and that test items have been systematically generated to cover the objectives. In Step 2, criterion validity is established through empirical means, and involves determining whether examinees who perform well on the test have really achieved the educational objective.

Using Classical Reliability and Validity. There has been considerable debate over the appropriateness of classical indexes of reliability and validity for CRTs. Some psychometricians have argued that since CRT items are selected to measure the achievement of specific educational objectives and not to discriminate among students, scores on CRTs can lack variation. This could arise in the following situation: Before instruction, none of the students have mastered the objectives, and they might all receive a score of zero on the criterion-referenced pretest, whereas after instruction, they might all receive very high scores on the criterion-referenced posttest. A lack of variation in student scores, it is claimed, would cause the traditional indexes of reliability and validity (that are based on variance) to be inappropriate (Popham and Husek, 1969).

Others have argued that when CRTs are adminstered to a heterogeneous sample that represents degrees of competence and receives different instruction on the objective, there will be sufficient variation in test performance to apply the classical statistical formulas (Klein, 1970, Harris, 1973). This latter stance is becoming the accepted view, and it is now held that the classical indexes can be estimated for CRTs using a heterogeneous population.

Theoretical Value of CRTs for Effectiveness Evaluation. Relying on the preceding theoretical discussion of the development and validation of CRTs, it is possible to ask whether, based on theoretical considerations alone, CRTs are appropriate to measure achievement for effectiveness evaulations and, in particular, for the SES. An effectiveness evaluation requires instruments that are reliable and valid, and that provide meaningful scores that can be used to make decisions about educational policy. In theory, there is an orderly set of developmental and validation procedures which, if followed properly, produces CRTs that are based on well defined sets of objectives and that can provide meaningful and useful score interpretations. Thus, from a theoretical perspective, CRTs are appropriate and desirable for measuring achievement in effectiveness evaluations, and in the SES. However, there are two important caveats that must be attached to this conclusion. First, there are persons who simply reject the notion of criterion-referenced testing and, with it, the meaningfulness of any CRT score interpretations. If an evaluation is being commissioned by individuals who share this view, then CRTs should not be used since the resulting information, although theoretically sound, is likely to be ignored. Second, as is the case with norm-referenced tests, not all CRTs provide the same type of score interpretation. Some CRTs are reported and interpreted in terms of the number of items passed per objective, and many educators and policymakers find this type of score interpretation, by itself, to be inadequate for most effectiveness evaluation purposes. However, a rejection of this type of score interpretation is not equivalent to a rejection of the notion of CRTs since it is just one of several acceptable ways of interpreting CRT scores, and since there is no reason why CRT scores cannot be supplemented by comparative data.

A REVIEW OF CURRENTLY AVAILABLE CRTS

To determine if currently available (1975) CRTs are technically sound, and if they have been designed to be easily used for evaluation purposes, a set of required CRT criteria was identified. Copies of currently available CRTs were obtained from publishers and were evaluated using the review criteria. Our intent was to determine the practical appropriateness of using CRTs for the SES, based on the results of the review.

Generating Review Criteria. To structure the review of CRTs, a set of criteria was generated. The criteria reflect the characteristics generally accepted as being necessary and appropriate for an effectiveness evaluation study. To obtain the criteria, we reviewed the literature, the unique needs of the SES, and criteria already developed and used for reviewing achievement tests.

Obtaining CRTs. A list of publishers of educational tests was compiled using test review books (Buros, 1965, 1972; Hoepfner et al., 1970; Hoepfner et al., 1971; Hoepfner et al., 1974), personal contacts, and library sources. It should be noted that publishers on the list were not necessarily known as marketers of CRTs because it was not always possible to predict in advance who published CRTs and who did not, and because it was considered important to include as many publishers as possible in the review.

A letter was sent to each publisher requesting that he send the following information about any criterion-referenced math or reading tests available:

- Detailed descriptions of the test battery at each grade
- Sample tests for reading and math at each grade
- Lists of objectives or domains for reading and math at each grade
- Directions for administering and scoring reading and math tests at each grade
- All technical manuals, field test reports, expert reviews, or test-analysis information
- Information about special features like scoring services or cassette-recording directions
- Cost information
- Name and title of person to be contacted for additional information

Some publishers responded to the inquiry by stating that although they did have math and/or reading tests available, they were not able to provide more than a brochure without a charge SDC decided not to purchase any materials. This decision was based on the expectation that the size and scope of the evaluation study would warrant an investment on the part of an interested publisher. Justification of SDC's decision appeared to follow since the publishers who did send materials usually had tests that were responsive to SDC's needs as described in its letter. Other publishers opted to lend SDC materials and requested that they be returned at the conclusion of this investigation. Only the 28 CRTs in reading and/or math that were accompanied by copies of the test(s) and test manuals were received and reviewed; the names of the publishers and the test systems are presented in Table 10-1. Each of the 28 CRTs was independently reviewed twice using the set of criteria generated for this purpose, and discrepancies were resolved by both reviewers. Any remaining questions, usually resulting from unclear or insufficient information, were followed up with phone calls to the publishers.

Explanation of Review Criteria

There are 18 criteria against which CRTs were reviewed. For this review, reading and language arts were considered to be one subject area, and math a second subject area. All

Table 10-1

Test Systems That Were Reviewed, by Their Publishers

Publisher
Name of Test System

American Guidance Service
- Key Math (Diagnostic Arithmetic Test)
- Woodcock Reading Mastery Tests, Form A

American Testing Company
- Mathematics Inventory Tests
- Reading Inventory Probe I

CTB/McGraw-Hill
- Comprehensive Tests of Basic Skills, Form S (CTBS/S)—Mathematics
- Comprehensive Tests of Basic Skills, Form S (CTBS/S)—Reading
- Diagnostic Mathematics Inventory
- ORBIT (Objectives-Referenced Bank of Items and Tests)
- Prescriptive Reading Inventory

Educational and Industrial Testing Service
- Tests of Achievement in Basic Skills (TABS)—Mathematics
- Tests of Achievement in Basic Skills (TABS)—Reading

Educational Development Corporation
- Individualized Criterion-Referenced Testing—Mathematics
- Individualized Criterion-Referenced Testing—Reading

Harcourt Brace Jovanovich, Inc. (The Psychological Corporation)
- Skills Monitoring System—Reading
- Stanford Mathematics Tests (1973)
- Stanford Reading Tests (1973)

Houghton-Mifflin
- Individual Pupil Monitoring Systems—Mathematics
- Individual Pupil Monitoring Systems—Reading

Instructional Objectives Exchange
- Objectives-Based Test Set—Mathematics
- Objectives-Based Test Set—Reading

National Evaluation Systems
- Comprehensive Achievement Monitoring (CAM) Maintenance Package—Mathematics
- Comprehensive Achievement Monitoring (CAM) Maintenance Package—Reading

Richard Zweig Associates, Inc.
- Fountain Valley Teacher Support System—Mathematics
- Fountain Valley Teacher Support System—Reading

Scholastic Testing Service
- Mathematics—Analysis of Skills
- Reading—Analysis of Skills

Science Research Associates
- Mastery: An Evaluation Tool, Mathematics
- Mastery: An Evaluation Tool, SOBAR, Reading

subtests or tests of individual objectives at the same level were grouped together and considered as a single reading or math test. Following is an explanation of the review criteria.

1. *Coverage of specific skills.* A test must cover skills in reading (language arts) and/or math since the federal mandates for the SES require measuring achievement in these areas. Examples of basic skills are reading comprehension, spelling, arithmetic, and telling time, as compared to tangential skills such as using the library or performing computations with a slide rule.

2. *Grade-level coverage.* Because of the scope of the SES, forms of the tests must be available for grades 1 through 9.

3. *Overlap of objectives.* Some or all of the test objectives must be measured at each grade level in order to permit comparisons across grade levels, or over time in terms of common educational objectives or skills. For this criterion, objectives or test items at different grade levels need not be worded identically, nor need a formal means of identifying them be provided.

4. *Number of test forms.* Due to constraints related to test administration and the time available for testing, there should be a limited number of test forms at each grade level. Just one test per grade level is preferred in order to avoid problems with reliability that can arise when several test forms are used.

5. *Directions for test administration.* A test should provide thorough and clear instructions for both the examiner and examinee. Directions concerning distributing tests, demonstrating sample questions, and administering should be provided in a detailed and easy-to-read form.

6. *Special equipment required.* Test administration should not involve any special equipment (such as cassettes or visual aids) aside from pencils and scratch paper.

7. *Time for testing.* A reading or math test should be designed to be completed within a class period. This usually involves no more than a maxmimum of 40 to 60 minutes, since CRTs have to fit efficiently within the time constraints of the planned information-collection activities.

8. *Group testing.* A test must be designed for group administration. This criterion was considered to be essential given the logistics of the information-collection strategy for the study which involves training local coordinators to train teachers who administer tests in group situations.

9. *Item-objective match.* Each test item should be coded to an objective or to the educational tasks and purposes the test claims to measure.

10. *Objective coverage.* There should be a sufficient number of items to measure each objective adequately. The number of items per objective should vary as a function of how broadly or narrowly an objective is stated and its level of difficulty.

11 *Objective scoring:* A test must use an objective scoring procedure since there are no plans to train individuals to interpret subjective scores.

12 *Machine scoring options:* The test must be available in or adaptable to a machine-scorable form.

13 *Score interpretation scheme:* A test must employ a criterion-referenced score interpretation scheme. Tests using CRT interpretations in addition to other types of score interpretation schemes were also acceptable.

14 *Curriculum dependence:* A test should not be based on the objectives of any particular curriculum or educational program.

15 *Costs of test per pupil:* The costs of testing pupils must be affordable for a large-scale study.

16 *Formal field test:* A test should provide documentation of field test activities. It is preferable that the field test participants be nationally and geographically representative, be a probability sample, and include sufficient numbers of minority and disadvantaged students to estimate bias.

17. *Information on item quality:* Information should be provided, based either on judgmental or response data, about item stability, sensitivity to instruction, sex/cultural bias, item-objective congruence, and equivalence.

18. *Information on test quality:* Information should be provided on test quality, based either on judgmental or response data, to include information about internal consistency, test stability, test-objective congruence, sex/cultural bias, sensitivity to instruction, and criterion validity.

Results of the Review

The results of the tests reviewed for this study are presented below. Because many of the 28 CRTs were intended for classroom rather than certification evaluation purposes, the review tended to make some CRTs look less excellent than they would have if they had been reviewed from another perspective.

1. *Coverage of specific skills:* Of the 28 tests reviewed, 15 were designed to assess only reading skills and 13 were designed to assess only math. All 28 tests focused on basic skills in reading and/or math, rather than on tangential skills, and thus met the criterion.

2. *Grade-level coverage:* Nine tests were available for grades K through 9, and thus met the criterion.

3. *Overlap or objectives:* Twelve tests appeared to measure the same objectives at all

grade levels. Sixteen tests appeared to have some overlapping objectives which were measured at most, but not all, grade levels, depending on the appropriateness of the objective and its level of specificity. It should be noted that to make common objectives, test publishers frequently used broadly-stated objectives or skill categories which they then translated into tasks and skills of varying complexity for different grade levels.

4. *Number of test forms.* Some CRTs had only one test form per grade level and others had as many as 31. Usually those CRTs that offered a limited number of test forms per grade level would include several objectives on a single form, while those featuring more test forms per grade level would assess one or only a few objectives per form. Three tests did not set limits on the number of tests that could be created from their bank of objectives and items.

5. *Directions for test administration.* Twenty-seven of the tests met the criterion by providing adequate directions both to the examiner and examinee for test administration.

6. *Special equipment required.* Twenty-six tests required no special equipment for test administration and, therefore, met the criterion. Two tests required the use of tape recorders or cassettes, and one test provided no equipment information.

7. *Time for testing.* Only two tests met this criterion. More tests (24) left time for testing open, but the reviewers judged that from their length they would require more than one hour of testing time. One CRT had no information about the time needed for testing.

8. *Group testing.* Twenty-five tests could be administered to groups and, therefore, met the criterion. Two tests were designed for individual administration only, and one did not provide information with regard to administration.

9. *Item-objective match.* Twenty-six tests had each item coded to an objective.

10. *Objective coverage.* The items tested for each objective ranged from one to 150. (It should be noted that the CRT with 150 items per objective was based on a computerized item bank from which tests of any length could be generated.) Tests with five or more items for each objective were judged to meet this criterion.

11. *Objective/subjective scoring.* Twenty-seven tests employed an objective scoring technique.

12. *Machine scoring option.* Eighteen tests met the criterion for machine scoring.

13. *Score interpretation scheme.* Twenty-seven tests met this criterion by using some type of criterion-referenced score interpretation scheme. Overwhelmingly, the scheme was expressed as an arbitrary master/non-mastery score, or the number of

items correct on a given objective. Of these same 27 tests, seven also employed norm-referenced interpretations.

14. *Curriculum dependence.* Twenty-two tests appeared to have total independence from a particular curriculum or instructional program. Six other tests also appeared to be rather general and independent, although they claimed to be based, in varying degrees, on a review of what is currently being taught in today's schools.

15. *Costs of tests per pupil.* Based on a purchase of all reading or math tests at the third-grade level, costs ranged from about five cents per student to $6.31 per student. Tests costing less than $1.00 per pupil per administration were judged to meet the cost criterion.

16. *Formal field test.* Eight tests provided documentation concerning field test activities. However, the information provided was remarkably sparse, with several exceptions. Those who did conduct field tests usually attempted to get some sort of geographic and national representation.

17. *Information on item quality.* Eleven tests reported having conducted item-quality studies based on both response data and/or expert review. Of these, attention typically was paid to item-objective congruence, item stability or equivalence, and sensitivity to instruction.

18. *Information on test quality.* Thirteen test publishers reported having conducted test quality studies based on response data and/or expert review. Of these, internal consistency, stability, test-objective congruence, sensitivity to instruction, and criterion validity (Step 1) were more frequently attended to.

Table 10-2 summarizes the results of the review for each test. Each reading or math test is identified by a numerical code, the codes are necessary because the publishers submitted their materials voluntarily and did not formally consent to a published review.

Practicality of CRTs for Effectiveness Evaluation

Relying on the preceding discussion of the characteristics of currently available CRTs, it is possible to ask whether, based on practical considerations alone, CRTs are appropriate for measuring achievement in the SES. The answer is 'no', from the review, it is clear that no CRT fully met the criteria. Further, the review not only uncovered serious practical problems that diminished the suitability of currently available CRTs for an effectiveness evaluation in general, but also brought to light additional problems which reduced the usefulness of CRTs for the SES in particular. These problems are summarized below.

1 *Many learning objectives.* Most of the CRTs reviewed had a large number of very specific learning objectives that were associated lessons. The reason for the use of many, narrowly defined objectives can probably be traced to the original use of CRTs

Table 10-2

Summary of Test Reviews

Test*	1. Coverage of Specific Skills	2. Grade Level Coverage	3. Overlap of Objectives Across Grade Levels	4. Number of Test Items	5. Complete Directions for Administration	6. No Special Equipment for Administration	7. Time for Testing Less Than 1 Hour	8. Group Testing	9. Item-Objective Match	10. Objective Coverage	11. Objective Scoring	12. Machine Scorable	13. Score Interpretations	14. Curriculum Independence	15. Cost Per Pupil Less Than $1.00	16. Formal Field Test	17. Information on Item Quality	18. Information on Test Quality
	Evaluation Criteria																	
1	X			X	X		X	X	X		X		X	X			X	X
2	X			X	X		X	X	X		X		X	X			X	X
3	X			X	X	X	X	X	X		X	X	X			X	X	X
4	X			X	X	X	X	X	X		X	X	X	X			X	X
5	X	X		X	X	X		X	X		X	X	X	X		X	X	X
6	X	X		X	X	X	X	X	X		X	X	X	X		X	X	X
7	X	X		X	X	X	X	X	X		X	X	X	X				X
8	X		X	X		X	X		X		X	X	X					
9	X	X		X	X	X	X	X	X		X	X	X	X	X	X	X	X
10	X	X		X	X	X	X	X	X		X	X	X	X	X	X	X	X
11	X			X	X	X	X	X	X		X	X	X					X
12	X			X	X	X	X	X	X		X	X	X					X
13	X	X	X	X	X	X	X		X	X	X		X	X	X	X	X	X
14	X		X	X	X	X	X		X	X			X	X	X			
15	X	X		X	X	X	X	X	X		X	X	X	X				
16	X	X		X	X	X	X	X	X		X	X	X	X				
17	X		X	X	X	X	X	X	X		X	X	X	X				
18	X		X	X	X	X	X	X	X	X	X	X	X	X				
19	X			X	X	X	X	X	X		X	X	X	X				
20	X			X	X	X	X	X	X		X	X	X	X				
21	X		X	X	X	X	X	X	X	X	X		X	X	X			
22	X			X	X	X	X	X	X	X	X		X	X	X			
23	X		X	X	X	X	X	X	X		X	X	X			X		
24	X		X	X	X	X	X	X	X		X	X	X			X		
25	X		X	X	X	X		X		X	X		X	X	X		X	
26	X		X	X	X	X	X	X		X	X	X	X	X	X			X
27	X		X	X	X	X	X	X	X	X	X		X	X			X	
28	X		X	X	X	X	X	X	X	X	X			X				

Note X = meets the criterion

*Tests are randomly ordered to preserve anonymity by teachers as an aide to individualizing and evaluating instruction. Nevertheless, an effectiveness evaluation of just one year of instruction at one grade level would generate information about an enormous number of objectives, thus complicating the management, analysis, and reporting of data.

2 *Numerous test forms* Many currently available CRTs provide separate test forms for each grade level that measures just one or a few objectives. For example, of the 28 tests reviewed, some had up to 31 separate forms for each grade level. The appearance of many test forms also probably reflects the original intention to use CRTs as classroom aides. In terms of national effectiveness evaluation, the logistics of administering a number of distinct tests complicates information collection activities, increases the chances of making errors, and increases the costs of conducting the evaluation.

3 *Time required for testing.* Most available CRTs take more than an hour of class time. The reviewers found that 23 of the 28 publishers claimed that their tests were untimed and thus left pacing to the discretion of the examiner, however, based on the number of test items, it is clear that one hour of test time is insufficient. In terms of the information-collection schedule planned for the SES, one class period of testing is the maximum time that can be devoted to a CRT.

It should be noted that some of the test publishers, recognizing time constraints, offered CRTs that had just one test item per objective. However, this is not a satisfactory solution since a reduction in the number of items will almost invariably bring with it a diminution in the test's ability to measure with precision any of the objectives.

4 *Matching CRT objectives with instruction:* Using CRTs in effectiveness evaluations that involve more than one educational program means determining relationships between objectives of the CRTs and the program so that achievement can be measured in terms of the objectives emphasized in instruction. However, obtaining this information is costly and complicated. Teachers can be asked, for example, to rate the objectives of the CRTs in terms of their relevance to classroom instruction, but teacher ratings can be unreliable. Instructional experts can be asked to analyze textbooks and curriculum guides; however, they cannot know for certain how these materials are being used in the classroom.

Another problem closely associated with that of relating CRT and instructional objectives concerns which objectives to test. Each student or classroom can be tested on just those objectives that are derived from the curriculum being used; or they can be tested on a large number of objectives that encompass all curricula; or they can be tested on a sample of objectives, some of which may be relevant to the curriculum, while the others may not. Depending upon the choice, the resulting evaluation information can be limiting in making comparisons or it can require considerable manipulation before interpretations can be made.

5. *Identifying common objectives:* A fifth problem with using CRTs in effectiveness

evaluation studies is that the same objectives are not always measured at all grade levels or, if they are, there is no system for identifying common objectives. Although the skills and content associated with an objective generally become more complex with increasing grade levels, to make comparisons over time or across grades it is necessary to identify skills or objectives that are related in terms of a conceptual framework or general content area. For example, in the fourth grade, a punctuation objective might focus on beginning sentences with capital letters and ending them with periods, while in the ninth grade, a punctuation objective might focus on the proper use of semicolons as alternatives to periods. Although both these objectives deal with the same skill area, grammar, unless they are formally referenced to that general skill area, the evaluator must make this instructional type of decision, one that is ordinarily not part of the evaluator's area of expertise.

6 *Validating CRTs* The procedures used to validate CRTs are not very sophisticated and field test results are not reported in any detail. When compared with the highly structured field tests conducted for norm-referenced tests, most CRTs are deficient with respect to the sample's size and representativeness, and/or the amount and precision of data presented in technical reports. It must be noted that test publishers have probably been reluctant to devote time and money to field testing because test theorists have not been able to provide them with an agreed-upon set of procedures for analyzing and reporting the CRT field test data. Assigning blame, however, is not the issue since the fact remains that a paucity of data is provided concerning the technical quality of tests and test items.

7 *CRT scores:* Most CRTs produce scores in one or two ways: either as the number of items correctly answered for each objective, or sometimes as mastery or non-mastery scores, where 'mastery' means correctly answering an arbitrarily selected number of items per objective. These types of score interpretations are accepted by theorists as a legitimate way of expressing CRT test scores, and such scores may have meaning for teachers who know their students and their curricula. However, for purposes of effectiveness evaluation, these types of interpretation alone are inadequate because they provide insufficient information for decision making and lose meaning outside the classroom.

8 *Financial considerations:* A final practical problem with using currently available CRTs for effectiveness evaluation purposes is that most are costly. This probably reflects the effort it takes to define domains; to develop the special features offered by CRTs, such as referencing the objectives to various school curricula; and to provide many short test forms so that they can be used efficiently for classroom instruction purposes.

CONCLUSIONS

There is no currently available CRT that is appropriate for use in the SES. This conclusion is based on practical, not theoretical, considerations. One major reason for the inappropriateness of available CRTs is that many of them have been designed for classroom, not evaluation, purposes, consequently, they are characterized by numerous, narrowly defined objectives, each measured on a separate test form. In the context of an effectiveness evaluation, these CRTs produce unwieldy amounts of information, require too much time for testing, and create logistical problems for test administration.

A second major practical failing of currently available CRTs is that field tests are either not documented or are performed inadequately. As a result, the reliability and validity of these CRTs is simply not known, and it is inappropriate to provide decision makers with information of unconfirmed quality. A third major failing of available CRTs is that the score interpretations are not as meaningful as the study demands. Most are presented as numbers of items passed, without criterion validity information or comparative data or supplements. Two additional practical failings include the costs of CRTs and the absence of mechanisms for tracking the same skills or objectives across grade levels.

A CRT that is feasible to use to measure achievement in an effectiveness evaluation such as the SES should be based on a limited set of objectives that represent essential competencies and basic skills, be proven reliable and valid, and be able to provide scores and are meaningful and useful. If the SES were to incorporate CRTs as part of the effectiveness evaluation in spite of the conclusions reached in this report, the following efforts would have to be undertaken prior to the administration of the test for purposes of interpretation:

- The goals and objectives of each classroom, curriculum, and instructional sequence of all participants in the study would have to be obtained. This report has briefly detailed the many practical problems this would involve.
- Some universal set of objectives would have to be judgmentally created. This set would then constitute the focus of the CRTs.
- One CRT system, preferably one that meets a maximum of the criteria defined previously, would have to be selected and, if possible, items sampled for each of the universal sets of objectives.
- The new CRTs would have to be field tested, both with students and with curriculum experts, to confirm that they meet the criteria.
- Score interpretations would have to be developed and validated both theoretically and empirically.

In essence, what would be called for is a test development and validation program far exceeding any so far completed for CRTs, rivaling those of the best standardized norm-referenced tests. Even then, the cost-effectiveness of such an option would be highly questionable.

219

CHAPTER 11. THE MEASURE OF GROWTH IN PRACTICAL ACHIEVEMENT

Mary L. Spencer, David W. Bessemer, Nicolas Fedan, and Bobby R. Offutt*

Upon completion of a review of theory and practice in the assessment of functional literacy and practical achievement, it was concluded that no existing instrument would be appropriate for the students in the Sustaining Effects Study (SES). Consequently, an instrument-development effort was undertaken that resulted in a reliable instrument tapping functional or practical reading and calculating that proved to be well-received by students. The reception is due to the illustrated item stems and to their relevance to the lives of young people in most walks of life in the U.S. The Practical Achievement Scale is to be administered in both the fall and spring to all students in the SES in grades 4 through 6.

Due to increasing concern regarding the use of standardized achievement tests with disadvantaged and minority students, it was felt that a measure of more lifelike, non-academic, or functional instances of literacy in children should be employed to provide an adjunct index of growth in reading and math for the SES. Social indexes of literacy have usually been based on years of schooling or on grade-equivalent scores derived from standardized reading achievement tests. Assessments of literacy based on such indexes have proved defective in two major ways. First, they apparently have led to a serious underestimation of the problem of literacy in the United States, and second, they have failed to provide a meaningful description of the actual capabilities of the population in terms suitable as a basis for policy. The defects of such indexes have been summarized by Harman (1970), Bormuth (1973), and Nafziger et al., (1975).

In recent years, an increasing emphasis has been placed on determining the actual abilities of adults performing practical reading tasks involved in real-life situations, particularly those having social and economic utility. Harris (1970, 1971) surveyed adult performance in filling out common application forms and answering questions based on newspaper employment advertisements. The National Assessment of Educational Progress (1971) conducted a nationwide survey of adult performances on a number of reading tasks based on practical materials. The Army has devised measures of reading, listening, and computa-

*The authors were affiliated with Pacific Consultants, Inc. of Berkeley, California, which was under subcontract to SDC for the development of a test of functional literacy. Several test authors, publishers, and agency sources of tests were most helpful in the preparation of the instrument. Their efforts to relay materials and information in an expeditious manner are appreciated, particularly in cases where materials pending copyright were entrusted to the reviewers. Mention is especially due to Dr. Norvell Northcutt, Adult Performance Level Test, Dr. Kenneth Majer, Basic Skills Reading Mastery Tests, Dr. Marilyn Lichtman, Reading/Everyday Activities in Life, Ms. Jan Algozine, New York State Basic Competency Tests, Dr. Harold Wilson, National Assessment of Educational Progress, and the Psychological Corporation, Fundamental Achievement Series.

tional capabilities required in specific military occupations based on job-related written materials (Sticht, et al., 1972).

In the course of developing tests designed specifically to measure adult functional literacy, substantial progress has been made in conceptualizing functional literacy and in the methodology of assessment. The work conducted by Educational Testing Service (ETS) has established a solid body of data on the reading activities of American adults (Murphy, 1973) A survey based on a national probability sample identified the materials commonly read during everyday general activities, the duration of reading, and the perceived importance of the reading activity. Based on the survey, a large number of performance tasks were developed. Each task was classified by type of material and by type of socioeconomic benefit provided by the task. An advisory panel evaluated the importance of the tasks, and a national survey of adult performances was conducted on 170 selected tasks. The data from the survey were later used to estimate the economic value of functional literacy skills (Murphy, 1975).

A domain-referenced test of functional literacy (R/EAL) was developed by Lichtman (1974). Through logical analysis, she identified nine common classes of reading materials representative of everyday life activities and selected one specific type of material within each class. A behavioral objective was written to cover the terminal and subordinate tasks in each hierarchy, using realistic facsimilies of actual sample materials.

DEVELOPMENT PROCESS FOR THE PRACTICAL ACHIEVEMENT SCALE

In service of the need to supplement indexes of reading and math achievement derived from standardized achievement tests, an extensive search and review of relevant literature was conducted. This process resulted in a dissatisfaction with the term 'functional literacy' because of its connotative limitation to reading written words because real-life activities include work with numbers and computations, the less restrictive term 'practical achievement' was substituted. The review process also resulted in the production of criteria for the evaluation and selection of tests of practical achievement for school children; the implementation of those cirteria on a set of candidate instruments; and the conclusion that a new test would have to be developed to meet the goals of the SES.

Definition of Practical Achievement. No definition of functional literacy has yet been widely accepted. Bormuth has suggested an all-inclusive definition in the following terms:

> In the broadest sense of the word, literacy is the ability to exhibit all of the behaviors a person needs in order to respond appropriately to all possible reading tasks.

Nafziger et al. (1975), point out that literacy, unlike reading, refers both to basic reading skills and to socially appropriate reading behavior. They indicate that functional literacy implies reading for a purpose, and a purpose related in some way to social utility. The U.S. Department of Education has been quoted by Nafziger et al. (1975), as defining a literate person in the following terms:

...one who has acquired the essential knowledge and skills in reading, writing, and computation required for effective functioning in society, and whose attainment makes it possible for him to develop new aptitudes and to participate actively in the life of his times.

None of these definitions offers very specific guidance to the operationalization of the functional literacy concept as a basis for development of an effective assessment device. A definition is offered below that lends itself to operationalization in terms of the wider dimensions of practical achievement.

Practical achievement in school children is the capability of performing in a consistently successful manner those reading and computation tasks that:

- are normally encountered in the course of everyday life activities by a majority of children in non-school settings;
- are normally encountered repeatedly, or involve a substantial duration of activity;
- involve commerce with particular types of materials commonly found in the environment of the child;
- involve non-technical language and symbolic representation;
- require specific observable behavior in relation to the material;
- serve definable types of social function; and
- are regarded as important by the child or established authority figures, or have demonstrable and non-trivial socioeconomic benefits.

Lacking a survey of reading activities for children comparable to that conducted with adults by ETS, it was relatively difficult to fully operationalize this definition of practical achievement. Nevertheless, progress toward this goal was accomplished through logical analysis, the advice of experts, and by capitalizing on certain aspects of previous work done with adults. Toward this goal, a panel of experts was convened to refine the definition and to provide guidance in the selection or construction of the needed test. This panel was composed of Dr. Marilyn Lichtman, Virginia Polytechnic Institute and State University; Dr. James Sanders, Western Michigan University; Dr. Thomas Sticht, Human Resources Research Organization, Western Division; and Dr. James Vasquez, Far West Regional Laboratory. The present definition and the decision to broaden our concern to that of practical achievement were direct results of a most rewarding interaction with the panel.

The first major point of the definition that was adopted is that tasks should be "normally encountered in the course of everyday life activities by a majority of children in non-school settings." This aspect was operationalized by seeking panel nominations for representative tasks and by informally surveying activities of a diverse sample of children, with special emphasis upon disadvantaged and minority children. As a result of this step, a list of life activities is presented in Table 11-1. In connection with each area of life, several significant entities are listed that could serve as effective stimuli-arousing associations to a reading or computation task.

A systematic classification of materials was developed, both as a stimulus to memory and as a refinement and explication of the third point in the definition, and is presented in Table 11-2. The table has been based largely on the classification of materials used by ETS (Murphy, 1973) and Lichtman (1974).

There is presently no basis for explicitly defining the extent of language and symbolic representations involved in chilren's practical achievement. Clearly, a child should not be expected to read language and handle concepts that are not yet incorporated into his spoken language competence, or that lie outside his normal realm of experience. One such restriction, confining the test to non-technical language, has been included in the definition. Language is considered technical if it occurs largely within a narrow field of social activity, if it is usually known primarily by persons directly engaged in that activity, and if

Table 11-1

Areas of Life Activity for a Test of Practical Achievement

1. Personal Maintenance
   a. Food
   b. Clothing
   c. Health
2. Personal Relations
   a. Family
   b Friends
   c Relatives
   d Pets
3. Institutional Relations
   a. School
   b Church
   c Club
   d Police and Fire
4. Neighborhood Locations
   a. Home
   b Shopping
   c. Postal
   d. School
   e. Medical
   f Recreation
   g. Library
4. Home Maintenance
   a. Appliances
   b. Yard and Garden
   c. Furniture
   d. Radio, TV, Stereo
6. Travel
   a. Bicycle
   b. Bus
   c Car
7 Leisure
   a. Sports
   b. Games
   c Toys
8. Communications Media
   a. Newspapers and Magazines
   b Comics
   c Television and Radio
   d. Books
   e. Movies
   f. Mail
   g Telephone

222

Table 11-2

Types of Materials for a Test of Practical Achievement

Signs, Labels
Schedules, Tables
Maps, Diagrams
Categorized Listings
Directions, Instructions
Advertisements, Announcements
Forms
Personal Communications
Instruments, Controls
Technical Documents
Discourse, Narrative

most children are unfamiliar with it. In the area of computation, no advanced forms of computation or symbols that children have not had an opporunity to learn should be included.

At this time, it is also difficult to propose a system of behavior categories that will provide a valid delineation of the specific behaviors involved in functional literacy tasks. To a large extent, the nature of the reading tasks seem implicit in, and inferrable from, the material associated with the tasks. For example, the typical behavior involved in using a telephone directory is describable in terms of a systematic search algorithm, given a specific entry name, proceeding through subgoals defined by a sequence of alphabetic cues, and resulting in location and retrieval of a specific numeric code from the listings. The usual behaviors associated with filling out forms include reading and comprehending headings or questions which identify requested information, retrieving personal information from memory or available records, and writing the information in appropriate blanks or checking off appropriate alternatives.

A systematic task-skills analysis of a considerable number of such tasks is required to define a comprehensive set of component skills covering the domain of practical reading tasks. Lacking such an analysis, an informal list of behavior categories is offered in Table 11-3, based on an examination of the materials used by ETS in constructing test items (Murphy, 1973) and those listed by Lichtman (1974). The kinds of material most usually found associated with each behavior category are also listed in the table.

A somewhat different tack can be taken in specifying the computational behaviors. Real-world computational problems can be distinguished on the basis of the computational operations, the numerical and measurement content, and the means of computation. This

Table 11-3

Preliminary Classification System for Behavior Exhibited in Functional-Literacy Tasks

Behavior	Material
Systematic Search	Categorical Listings, Advertisements
Choice Discrimination	Signs, Labels, Maps, Diagrams, Instruments, Controls
Preference Selection	Categorized Listings, Advertisements
Retrieval of Personal Data	Forms
Selection and Storage of Information	Personal Communication, Discourse, Narrative
Performing Sequence of Operations	Directions, Instructions, Maps, Diagrams
Contingency Identification	Signs, Labels, Directions, Instructions, Technical Documents
Accuracy Verification	Technical Documents, Forms
Comprehension of Information	Personal Communication, Maps, Diagrams

provides the basis for a facet design of computational categories shown in Table 11-4. As shown, the child will perform computational tasks associated with real-world materials requiring operations, based on signed numbers representing physical quantities, expressed in units of measurement by appropriate means.

ETS identified eight categories of socioeconomic benefits as shown in Table 11-5. These seem to overlap to some extent with the catagories of life activity presented in Table 11-1.

Table 11-4

Facet Design for Computational Categories

Operations	Signs	Numbers	Quantities	Units	Means
Addition	Positive	Integers	Length	English	Mental
Subtraction	Negative	Fractions	Area	Metric	Paper and Pencil.
Multiplication		Decimals	Volume	Kitchen	
Division		Percentages	Weight		
			Time		
			Dry Measure		
			Liquid Measure		
			Speed		
			Money		
			Cost/Unit		
			Angular		

Table 11-5

Functions of Literacy Behavior

Economic
Occupational
Education/Culture
Recreation
Health
Maintenance
Personal Relationships
Citizenship

Requirements for a Measure of Practical Achievement. A general description of the characteristics desired in a test of practical achievement is described. First, the instrument must clearly measure the operational definition of practical achievement that was developed for the study. Accordingly, practical achievement is viewed as the reading and computational skills needed by children as they deal with the contemporary non-school-related world. The test must be independent in the sense that it is specifically designed to measure practical achievement rather than being merely the reading or computational portion of some achievement test battery. The level, range, and content of the test must, furthermore, be appropriate for school children in grades 4 through 8 from disadvantaged backgrounds.

Some of the more practical requirements are that the cost of the test should be within the normal range of costs for comparable tests, it must be capable of being group-administered by non-expert school personnnel employing uniform procedures across the country, and it should be machine scorable.

If the test is a norm-referenced test, the norms pertaining to the population of the study should be available. If the test is criterion-referenced, the criteria on which the test is developed should relate to the study's definition of practical achievement and the criteria for score interpretation should be clearly specified. Technically, the test should exhibit adequate reliability and validity indexes.

Criteria for Selecting a Test

Characteristics of the test, the nature of the examinees, and the purpose of testing are important factors in selecting the practical achievement test. The criteria for test selection are based largely on the general guidelines provided by the American Psychological Association's Standards for Educational and Psychological Tests (1974), and the criteria employed in the CSE Elementary School Test Evaluations (Hoepfner et al., 1976). Additional criteria were suggested by review of reading and literacy tests for a Right-to-Read Evaluation, and the examination of tests of adult functional literacy performed at the Northwest Regional Educational Laboratory (Nafziger et al., 1975). The criteria define factors relevant to test selection, but were not concerned specifically with the measurement of practical achievement in grades 4 through 8 for the purpose of program evaluation. A number of general recommendations were not suitable in meeting the special requirements of the Title I study and were therefore modified as necessary. The proposed selection criteria are organized according to the six general areas: (1) test background, (2) psychometric quality, (3) appropriateness, (4) norm standards, (5) administration, and (6) interpretation.

Test Background. The purpose of the test should be explicitly stated, and examples should be mentioned. The purpose statement should be clear to those individuals likely to administer the test. The construction of the test should follow closely the purpose for which the test was built. Thus, a diagnostic test should state how the test's purpose translates or agrees with the scope of tasks and operations to be covered. Such scope should be limited, well defined, and detailed. The test should indicate whether differences among minority groups were considered during test construction. If such consideration were exercised, items would have been sampled that depict the actual behaviors of students in these groups in extracurricular life activities.

Psychometric Quality. Criteria addressed in this section pertain to the validity, the reliability, the comparability of test scores, and the quality of normative standards. Factors contributing to the credibility of a test as measuring practical achievement are considered in terms of content, empirical, and construct validity. In terms of content validity, it is desirable that the test be representative of a definable population of items and performances with specific reference to the domain of practical achievement. The test should be specifically designed as a test of practical achievement. Disagreement on the validity of content will surely arise if the test was originally designed for some other purpose and if no explicit basis exists for judging the relevance of items. The stimulus materials should be representative of those commonly encountered in real-life reading and computational tasks. Confidence in the representativeness of materials would be increased if a population of such materials were defined, and if the composition of the population were described in terms of types or characteristics of materials and formed part of the definition of practical achievement used as the basis of test development. The performance required in the items should be as close an approximation as possible of the tasks and skills commonly required in real-life reading and computational performances of children.

Explicit classification and/or description of a domain of behaviors is desirable as part of the definition used as a basis for test development. The language and other symbolic representations should be representative of the symbolic content commonly encountered in real-life reading and computational tasks. Specification of the symbolic content in linguistic and mathematical terms can further strengthen and clarify the definition of practical achievement beyond the material and behavior specifications usually considered. Such specifications could be particularly helpful in defining levels or ranges of competence in relation to the domains of materials and tasks. The materials and tasks should be representative of the socioeconomic functions commonly encountered in real-life reading and computational tasks. A classification or description of socioeconomic functions, and the benefits or values of performance, should be part of the definition of practical achievement that is used as a basis of test development. Such a classification or appraisal system would help insure that functionally significant, rather than trivial, performances are represented. The materials and tasks of the functional literacy test should not be referenced to specific program objectives; the test is intended to provide an objective criterion by means of which the effectiveness of various programs can be judged in the area of functional literacy. The definition of practical achievement should be operationalized by the specification of a set of criterial tasks referenced directly to the characteristics of materials, behavior, symbolic content, and functions employed.

In terms of empirical validity it is desirable that there be empirical evidence relating the test scores meaningfully to other variables. It is advantageous that the test has been correlated in previous studies with other measures taken at the same time. The number and quality of studies, the number of variables, and the diversity of variables all contribute to the evidence bearing on the meaning of a given literacy score. It is also advantageous that the test be correlated with measures taken at some later time. Studies should also relate the test to important psychological, educational, or socioeconomic variables. All these empirical relationships should be interpretable in terms of prevailing educational, psychological, and socioeconomic theory. The measure of practical achievement should relate sensibly to variables that reflect components of functional literacy and to variables that are independent of functional literacy. Finally, the test should be sensitive to the effects of appropriate independent variables so that there is some assurance that effects will be revealed in the present study as well.

Considerations of construct validity have to do with the theoretical basis of the practical achievement concept. They are of lesser importance in judging validity than are content and empirical criteria, given the practical concerns of this study. The conceptualization, development, and empirical validation of the test should be grounded on relevant psychological, linguistic, and educational theory. Particularly important in this respect is the availability of a task-skills analysis which would define the components of practical achievement, indicate hierarchical relations among components, and the relationship of performance to basic cognitive information processing operations. Such a theoretical foundation is useful in generating hypotheses and interpreting results. Such formulations would provide a basis for tying changes in practical achievement to specific educational practices and would provide a basis for interpretations of findings concerning relevant socioeconomic variables, and the function and benefits of literacy.

The selection should also consider the test's reliability. If alternative forms are available, they should be based on parallel items with comparable item statistics. The forms should correlate .80 or above at every grade level in the fourth- to eighth-grade range. Test-retest correlations should be .80 or above over brief time intervals, i.e., one month or less. Reliability coefficients could be lower over longer intervals, particularly when instructional experiences have intervened, having a substantial effect on the level of functional-literacy performance. High internal consistency is not a necessary criterion for the test, since a test which is highly homogeneous is not likely to represent the full diversity of tasks that should be sampled for practical achievement. A statistic that allows an interpretation of the reliability of each score is desirable. If the test discriminates at various age levels, the standard error of measurement for each level would show how well this differentiation is accomplished. Finally, the procedures used in item sampling should be clearly defined and replicable. It is necessary that test information indicate the relevance and representativeness of the item pool in relation to the aspects specified in the definition of practical achievement Procedures used in selecting items from a pool for inclusion in the final test should be based on observations of actual behavior. Items should include a wide range of difficulties, including some items relatively easy for fourth-grade children, and some items relatively difficult for eighth graders.

Appropriateness. The third set of criteria concerns the appropriateness of the test for the examinees. They focus on the areas of instructions, on items and format, and on procedure, and ensure that irrelevant sources of difficulty are eliminated from the test.

The instructions should be appropriate in orientation and tone, inoffensive in content, and comprehensive, with vocabulary and syntax suitable for children in the fourth- to eighth-grade range They should provide an honest explanation of the purpose and intended use of the test Further, they should precisely and completely describe all requirements of the tasks presented in the items so that the examinee has all the information needed to adopt an effective performance strategy. Sample items should be included that accurately illustrate task requirements and the level of difficulty of the tasks. A standardized script should be available for fluid oral reading by non-expert examiners.

The items should be relevant, up to date, and interesting for children in the fourth- to eighth-grade range, so as to arouse intrinsic motivation in task performance without extensive exhortations being required to induce cooperation and effort. The content of the items should not involve any invasion of privacy, or any sexist, racist, or otherwise offensive aspects.

In terms of physical quality, the paper should be of good quality, the print bold and readable, and the illustrations clear, up to date, and preferably in color. The test should be effectively arranged and cued to facilitate recognition of items as units, the perception of the relation of item stems to answers and examinee responses, and the progression of successive items and pages. Most examinees should have sufficient time to attempt most items. Sectioning of the test, with timing instructions for each section, may help to maintain appropriate pacing in the brief time allotted. The responses should be marked in a fashion permitting machine scoring. Each item should require one simple and direct response, with no multiple steps or complications other than those intrinsic to the task represented by the item. Several items might be used, based on the same stimulus materials, provided that the relationship of each item to the stimulus is clear.

Norm Standards. Norm data are not essential in view of the large sample to be tested in the SES and the emphasis on program evaluation. It is desirable, however, that some norm information be available for the fourth- to eighth-grade range that is representative of racial, ethnic, sexual, geographic, and socioeconomic strata. It is desirable that such data be reported separately as well as combined over the strata represented in the sample. It would also be useful if item statistics were reported both for the whole sample and for the separate strata.

Administration. Non-expert school personnel should be able to administer the test with very little training. The services of a specialist or a testing expert, or extensive training, should not be required. The test should require no more than 30 minutes of testing time. Tests taking longer than 30 minutes should be easily modifiable for a shorter length, with no more than normally expected loss of reliability. Test administration in usual classroom settings, to group sizes in the normal range for intact classroom groups, and without the necessity of special equipment, is very desirable.

The test materials should be entirely of the paper-and-pencil test variety, with no special manipulanda, slides, or other unusual components; the costs should be in the normal range of paper-and-pencil tests.

Interpretation. A high quality test manual should be available, describing the test and how to interpret its scores. The test scores should be highly meaningful and understandable by a non-technical audience including the general public. It is desirable that the implications of given test scores for educational practice or public policy be clear and relatively direct.

Evaluation of Existing Tests

Six instruments were selected for review according to the criteria. They were selected because they possessed some property or set of properties that placed them within the range of promising instruments for the study's purposes. It should be clear from the outset that the judgments made of these tests relate only to the potential usability of the instruments in the SES, and the test evaluations should not be construed as either indictments or recommendations of the instruments for adoption in other contexts. Summary evaluations of each test are provided below.

Adult Performance Level Test of Adult Functional Competency (APL). Although the APL measures both reading and computational skills, it is intended for use with an adult population. In the reviewer's judgment, most APL tasks are much too diffficult for children.

From a test-construction point of view, although efforts were made to select tasks actually encountered by adults, there was no evidence that observations of adult behavior were made so as to verify these tasks. The absence of validity data does not allow a judgment on the relationship between the APL and other tests and variables important to the concept of practical achievement.

Modification of this test would depend on the availability of easier items. Since the APL was constructed for use with adults, it is unlikely that enough easy items can be obtained from the item pool to construct an instrument appropriate to the study sample.

Basic Skills Reading Mastery Test (BSRMT). Two equally large problems exist with the

BSRMT from the standpoint of use for the SES. no computational skills are included and the test was constructed to be used with a population aged 12 to 18 years old. From a conceptual point of view there are also problems, the most important of which is that the manner in which stimulus materials were obtained for item construction did not include actual observation of the people for whom it is intended. Modification of this test would hinge on the availability of an item pool from which easier items could be obtained. In addition, a computational subtest would have to be either constructed or adapted from some other source.

Fundamental Achievement Series (FAS). The FAS, as is, cannot be rated appropriate for the SES, primarily because it was intended for, tested, and normed with grades 6, 8, 10, 12 and with various industrial groups. Additionally, there is no information on item construction, and thus it is impossible to determine whether observations of reading behaviors were used for item generation. On the positive side, this test includes both computational and reading skills, and it has been widely normed. Modification of the FAS to suit the purposes of the study could probably be accomplished, provided that the publisher has item statistics on the remainder of the item pool.

National Assessment of Educational Progress in Reading-Released Exercises (NAEPR). The advantages of the NAEPR exercises are that they have been used with children of the age group to be involved in the SES and item statistics are available for several examinee variables. Further, they could be formatted in a test package of appropriate length and to meet other criteria of administration.

In spite of the positive qualities, the NAEPR exercises have several undesirable features. The most glaring rises from the lack of a computational subtest. In addition, the stimulus materials are not judged to be intrinsically motivating and post-pilot work with them has revealed them to be sufficiently difficult for children in the fourth- to eighth-grade range as to generate some degree of examinee resistance. The item development process has the conceptual difficulty of having been created by experts, rather than having flowed from observations of real-life experiences of the population of potential examinees.

The possible utility of the NAEPR exercises to the Title I study is two-fold. First, an examination of the item statistics of certain items judged to have qualities of intrinsic motivation and suitability may result in the selection of particular items. Second, the NAEPR items may have heuristic value to the development or modification of a practical achievement assessment tool. As they stand, however, the collection of items represents an unacceptable means of assessment for the SES. At a minimum, this collection would have to be modified and supplemented extensively.

New York Basic Competency Test in Reading (NYBCTR). The lack of relevance of the items of the NYBCTR to children having the relevant age and socioeconomic attributes presents a serious drawback for the test's consideration. Futher, the test does not contain a computational section. The issue of item suitability is so problematic that it renders the modification of this instrument impractical.

Reading/Everyday Activities in Life (R/EAL). The main strength of the R/EAL is the real-life characteristics of its item formats. Beyond this, little can be said to recommend it for use. There is no evidence that the items can be used with children in grades 4 to 8. No computational subtest is available. The use of cassettes or tapes makes it unusable for present purposes, although the newly prepared instruction script could alleviate this drawback. A passing grade of 80 percent is the only normative data presented, and is not related to other kinds of real-world performances.

Implications and Recommendations From the Test Review

As a result of the literature review, and the test-evaluation activities, a series of conclusions and recommendations emerged. None of the reviewed tests satisfactorily meets all criteria. The inadequacies of the tests for purposes of the SES are basic. No test could be found that is appropriate to the fourth- to eighth-grade age group. All six of the tests contain various materials that are commonly encountered in real-life reading, with test items constructed from these materials. However, none of the materials was obtained by actual observation of the behavior of children. Two of the tests measured both reading and computational skills but cannot be used for the Title I study without modification because they are too difficult. One could probably be modified at less expense, since it already covers the higher end of the intended sample, i.e., grades 6 and 8. Instructions, however, are tape recorded, and would therefore require modification. Although there is no evidence that the test items in either were built around actually observed, real-life reading behaviors, the APL items are accurate facsimilies of literacy stimuli commonly encountered by adults. Various other important criteria were lacking in these tests, but since the most basic criteria were not met, it is a moot point to discuss additional inadequacies.

The remaining tests, which measured reading skills only, have various advantages and disadvantages. NAEPR had the most information on each item. Although individual items rather than an administration-ready test package are actually available, if the construction of a test is envisioned, then serious consideration should be given to some of the items presented. Item statistics are available by sex, race, geographic region, size and type of community, and age.

The favorable aspect of the BSRMT is that it has a test form for children 12 years of age. The lack of realistic facsimiles as item materials depreciates its value to the disadvantaged population of children to be studied. Modification of this test would require that a set of easier items be added for 9- to 11-year-old children and that the representativeness of stimulus materials be improved. Further, test instructions would have to be converted to an oral presentation mode.

The R/EAL has the important feature of presenting its items as actual photographs, or true-to-life drawings of the objects that contain the reading matter. This is an extremely desirable characteristic of item presentation, especially where minority groups are to be tested. Unfortunately, most of the items were constructed for an intended population of high school graduates and older. Modification of the R/EAL for use in the study would succeed only if certain items were selected which are judged to have relevance to fourth- to eighth-grade children, and then supplemented with other easy and appropriate tasks. The instruction script and other details of administration would require minor modifications in order to make them entirely compatible with the testing purposes of the SES.

No test exists that fully meets the needs of the study. It would be possible, however, to construct a new test from multiple sources. This procedure would entail the combination of a computational section with either a partially suitable test of reading competency, or with a reading section constructed of items drawn from several instruments. In either case, some set of easy items would have to be added to both the reading and computation portions of the assessment tool. These easy items could be either new creations or, more likely, modifications of items extant in some of the six tests reviewed. Another version of this option would consist of constructing a test from all possible sources, sampling and building items Any test produced by these methods would then have to be pretested, and modified at least once before it would be ready for a field test. This approach unfortunately precludes the advantages of generating literacy tasks from the actual experiences of disadvantaged children. For this reason, it is concluded that a wholly new test must be developed.

The development of a new test would entail sampling the actual reading behaviors of children who match the age and demographic characteristics of those participating in the SES. A technically sound item-building phase would then be required, with careful pretesting of the final instrument. The advantage of this option is that the final instrument would be a high quality test that would be suitable for a wide range of future applications. Moreover, it would be most responsive to concerns regarding the testing of disadvantaged children.

THE PRACTICAL ACHIEVEMENT SCALE

Efforts in the development of a custom-tailored test of practical achievement were based upon the desirable features and aspects of the tests of functional literacy reviewed, but focused upon remedying those aspects that made them inappropriate for use in the SES.

The measure of practical achievement in school children was prepared in a manner consistent with the test administration requirements specified in the criteria for test selection. Administration criteria are as follows:

Personnel. Non-expert school personnel should be capable of administering the test with very little training. The services of a specialist or a testing expert, or extensive training should not be required.

Scheduling. The test should require no more than 40 minutes of testing time (preferably 20 minutes) on one occasion of testing. Tests taking longer than 40 minutes should be easily modifiable for a shorter length, with no more than the normally expected loss of reliability.

Setting. The test should be capable of administration in usual classroom settings, to group sizes in the normal range for intact classroom groups, and without the necessity of special equipment.

Scoring Method. The test should be scored in an objective manner by machine. Machine scoring should be highly fail-safe and reliable, without complex error checking routines to proof the results.

Materials. The test materials should be entirely of the paper-and-pencil test variety, and require no special equipment or materials.

Clinical Pretest

In initial test construction, input was solicited from teachers, parents and other social or educational experts. This provided information on the construct validity of the scale. Judgments of item difficulty were aided by the results of an educational reading project that used stimulus materials and items from the Practical Achievement Scale (PAS). This study was conducted by a University of Califonia graduate student, who made available to us some of his preliminary results. The data showed that for the age group used in the study (12 to 15), most of the items were too easy since the proportion of children answering most items correctly was nearly 95 percent. These data suggested that the items might be too easy for the target population of the SES. Consequently, several easy items were deleted and more difficult ones were added.

A clinical pretest was performed in which the instrument was administered to nine children in the requisite 9- to 13-year-old range. In the clinical pretest, verbal comments were sought on each item of each stimulus object. On the basis of these comments, items were deleted or revamped to enhance clarity and avoid irrelevant or misleading wording or wording that was too difficult. The quality of each item was also tested by ensuring that it called unambiguously for the skills that it was intended to tap rather than involving some mental 'trick.'

Based on all the information gathered or made available to us, a further item revision was carried out, yielding the field-test version of 49 items and 23 stimulus materials. The easy questions required less than a minute and the more difficult questions seldom required more than a minute. Approximately 30 to 40 minutes are needed by most children to finish the entire set of exercises, with the test administrator reading the questions and optional responses aloud.

The Field Test

The new version of the PAS would clearly have to be field tested before being employed in the SES. The field test would provide information on the appropriateness and effectiveness of items, on time and pacing, and on student reaction to the test. Based upon consideration of the life experiences of the country's major minority-group children, 23 item bases with 49 items were selected from a larger pool of potential item bases that had been developed and pilot-tested by the Pacific Consultants test-development team. The item bases were all pictorial stimuli that presented information to be used in the solution of five alternative item questions.

The item bases and questions were thought to be appropriate and interesting to the disadvantaged and minority students as well as others, and were expected to provide a fairly good range of item difficulties, so that a final test could be selected on the basis of field-test data that would be generally easy for most students.

The field-test version of the Practical Achievement Scale was provided with simple instructions that indicated to the students what was to be expected of them and how they should go about answering the questions in the test. In addition, the "I don't know" response alternative was explained as a response of last resort. The instructions and the item questions with the alternatives were read to the examinees, so that the effects of academic-like reading achievement would be minimized. The field-test version of the test was printed to appear very much like an optically-scanable test instrument, which is the format to be taken by the final form of the test. In this manner, any problems with format would be uncovered during the field test.

The Field-Test Goals and Plan. The PAS was field tested at 15 schools across the country that were selected according to the four stratification types: (1) Northeastern, poor, urban; (2) Southeastern, poor, non-urban; (3) Northwestern, non-poor, urban; and (4) Southwestern, non-poor, non-urban. Two classrooms at grades 4 and 6 at each of the schools were administered the test under the standardized field-test procedures. A total of 750 fourth-grade students and 758 sixth-grade students completed the instrument in a valid manner.

The main purpose of the field test was to provide information for instrument revision. A second purpose was to provide some provisional statistical information about the final form of the scale. The questions that were posed for the field test are listed below:

1. Which items (pages) can be eliminated to reduce the length of the test to about 12 to 16 pages? Three criteria were employed in the process of eliminating items:

   a. Are item difficulty indexes appropriate? Items too difficult for sixth graders or too simple for fourth graders would be nominees for deletion from the final form.

   b. How closely is each item score (page score) related to the total score? Items and pages among those less correlated with the total score would also be candidates for elimination.

   c. Do the item scores (page scores) discriminate between the fourth-grade and sixth-grade students? Items that fourth-grade students do better on than sixth-grade students would be considered for elimination, as well as items for which there is no notable difference between the two grades.

2. How reliable is the total score for the reduced test (after undesirable items have been eliminated)? If the reliability is too low (less than .70), the least deletable items may be brought into the final instrument to increase the reliability.

3. Are the distractors of the remaining items functioning adequately? If some distractors (wrong alternatives) are so obviously wrong that too few students select them, or if too many students choose them, they will be replaced with new alternatives (but the replacements will not be further field tested).

In addition to the analysis of response data from the field test, trained observers were present in the classrooms at most test administrations to make formal observations of problems. The observers were to obtain information on the length of time it took to administer the test, problems that children had in following the instructions, problems with item pacing, problems with student interest or morale with items, and any overall impressions that might lead to the improvement of the test.

Analysis of the Field-Test Data. The data collected from the field test were analyzed by computer and by manual tabulation of information not amenable to computerized analysis procedures. The analyses are presented below, in the same order as the goals of the field test.

Are item difficulty indexes appropriate? Table 11-6, columns 2 and 3, present the percentages of students answering each item correctly, at both the fourth and sixth grades. Columns 4 and 5 present the item biserial correlations with the total test score for grades 4 and 6, in response to the question of how closely related each is to the total. A comparison of item difficulties between fourth and sixth grades can be made by comparing the difficulty indexes in columns 2 and 3 of Table 11-6. Based on the findings in Table 11-6, 19 items from 10 pages were deleted from the instrument to provide a test with 14 pages of 30 items. The retained items were then submitted to a program for the computation of the internal-consistency reliability of the reduced-length test. The reliability coefficient obtained is equivalent to that obtained by the Kuder-Richardson-20 formula. The coefficient for fourth-grade students was .89, and that for the sixth-grade students was .90. These estimates of test reliability were considered satisfactory, so the item-selection procedure was not iterated to obtain a larger coefficient. The response frequencies to each of the items of the reduced-length test were then examined, and minor changes in four of the alternatives were made. These changes can be expected to result in higher reliability for the instrument in late testing, and were made for this purpose. The reduced test total scores at the fourth grade correlated .84 with the standardized reading total (N = 315) and .69 (N = 331) with the standardized math total. At the sixth grade, the correlations were .81 (N = 307) and .74 (N = 327), respectively.

Results of the Field Test Analysis. The reduced-length test had a total of 30 items on 14 pages, as a result of the field-test analysis. Analysis of the observations of the test administration resulted in the following recommendations, which were incorporated into the present form of the PAS and into the administrative procedures for the operational years of the study.

Table 11-6

Item Difficulties and Item-Total Correlations From the Field Test of the Practical Achievement Scale

Item	Percentage Correct Responses		Total Test Biserial Correlations	
	Grade 4	Grade 6	Grade 4	Grade 6
1	94	95	.41	.68
2	72	87	.59	.43
3	51	70	.63	.74
4	23	47	.51	.62
5	86	94	.70	.76
6	39	54	.47	.52
7	86	92	.67	.68
8	67	86	.68	.87
9	49	79	.66	.82
10	72	85	.71	66
11	69	74	.18	.35
12	63	75	.52	.68
13	34	67	73	.83
14	37	61	.81	.75
15	40	61	.64	.61
16	27	55	.68	.77
17	66	83	.68	.77
18	70	86	.70	.78
19	43	53	.49	.43
20	51	68	.65	.72
21	69	83	.57	.59
22	11	20	.24	.46
23	53	60	.28	.24
24	67	80	.50	.57
25	39	64	.63	.65
26	60	81	.67	.77
27	57	76	.58	.67
28	53	74	.62	.70
29	49	70	.63	.74
30	71	87	.65	.83
31	66	83	.71	.78
32	34	56	.66	.62
33	80	89	.70	.85
34	36	44	.38	.35
35	30	54	.43	.55
36	48	68	.39	.49
37	45	70	.67	.77
38	43	65	.71	.67
39	40	60	.45	.65
40	43	57	.30	.45
41	38	58	.52	.48
42	61	85	.57	.69
43	41	63	.74	.69
44	46	72	.71	.77
45	34	57	.72	.76
46	36	58	.62	.62
47	19	27	.24	.35
48	54	82	.56	.61
49	19	44	.49	.68

236

1. The easiest item was moved to the instruction page and incorporated into the test directions as an example.

2. The test directions were printed on the cover page of the test.

3. Items with difficult-to-read symbols or abbreviations were altered so that the test administrator would have no trouble reading aloud.

4. Alternatives were not to be read to the students; only item stems were to be read.

CHAPTER 12. MEASURES OF AFFECTIVE GROWTH

Ralph J. Melaragno and Ralph Hoepfner

Available measures of aspects of affective growth in children were reviewed with the intent of selecting a set of scales as outcome measures for the Sustaining Effects Study (SES). The Survey of Student Attitudes was selected to assess attitudes toward reading and math. Additional items were generated in a similar format to assess attitudes toward school in general. These three dimensions of attitude appeared in the field test to be relatively independently and reliably measured constructs. A primary version of the scales was administered in the fall and spring to students in grades 2 and 3, but only in the spring to students in grade 1. An intermediate form was administered in the fall and spring to students in grade 4 and above.

The primary thrust of all compensatory-education (CE) programs is to improve the cognitive development of participating students. A parallel, but subordinate, objective is the improvement of students' school-related, non-cognitive development. For example, the California guidelines for Title I state:

> The main goal of every Title I project shall be to increase the academic achievement level of all eligible project participants to reflect a normal range and distribution of academic achievement for the target population as compared with the general population. Enhancing pupil self-image, motivating the pupil to achieve, improving his (sic) health, and raising his (sic) aspiration levels are to be considered supportive objectives that must be attained in order to meet the project goal.

Thus, while cognitive growth is the basic goal, non-cognitive improvement is viewed as an intervening variable—a precondition that enhances the likelihood of achieving cognitive growth, and that is influenced in turn by the success or failure of achievement in the cognitive area.

School personnel involved in compensatory programs typically express great interest in the area that is broadly termed 'attitude toward school' and that includes a student's feelings about peers, teachers, instruction, and learning. Most CE program descriptions include a concern with the improvement in students' liking for different aspects of school work, particularly liking for activities in reading and math (which are the main areas of concern in CE).

Students' affective behavior was to be measured as part of the SES. The longitudinal nature of the study provides an opportunity to measure changes in affect that occur as students experience the cumulative effects of CE. We are concerned with measuring the extent to which a student's affect is influenced by participation in a CE program, and the differential effects on affect of programs with identifiably different characteristics. While direct assessment of affective development would be preferred (e.g., through observation of students), this is precluded by the size of the sample, the burden that would be imposed on students, and funding constraints. Thus, the measurement task must be accomplished less directly, through self-report instruments completed by students in the study.

SELECTION OR DEVELOPMENT OF INSTRUMENTS

There are two methods of obtaining instruments for measuring affective behavior. One is to use existing instruments, and the other is to develop new instruments, either by creating new items or by selecting items from other scales. Each method has been used in earlier studies of CE and each method is accompanied by serious problems.

In its evaluation of Follow Through, Stanford Research Institute (1970) used a variety of existing instruments. Generally, the results pointed to limitations in the discriminating power of the instruments and a lack of strong relationships with program characteristics. Both the Equality of Educational Opportunity study (Coleman et al., 1966) and the Emergency School Aid Act study (Coulson et al., 1975) employed the technique of embedding affective items within a questionnaire completed by students. In both cases, the scales formed by the items were not successful in distinguishing between programs with different characteristics.

In a more narrowly focused study, the Educational Testing Service evaluated Compensatory Reading Programs (ETS, 1975) and, having found none of the available instruments adequate to the study, developed inventories to measure students' attitudes toward reading. The ETS instrument yielded unclear results in the evaluation of the different reading programs.

The major problem with using existing instruments has been that they may not be appropriate to the needs of a given study. Typically such instruments were developed for particular respondents and specific purposes, and their utility for other kinds of respondents is likely to be limited. The major problem in developing new instruments for a particular evaluation effort is the time and expense involved in determining the reliability, validity, and other psychometric properties of test items.

Criteria for Selecting Affective Instruments. In investigating the feasibility of using existing instruments or developing new ones for the SES, we first identified areas within the affective domain that appeared relevant to the study. We looked for areas of concern to an evaluation of CE, areas that have been measured in previous large-scale evaluation studies, and areas that current research has shown to be related to students' experiences in school. Application of these criteria resulted in the selection of 'attitude toward school,' 'self-concept,' 'locus of control,' and 'achievement motivation' as the areas with the greatest relevance.

Next, eight criteria for evaluating existing instruments were developed and applied to all instruments that had potential for use in the study. In an initial screening of possible instruments, standard references on affective measures and the files at the Center for the Study of Evaluation at UCLA were examined, and the following criteria were applied to all instruments:

- *Validity and Reliability.* There should be information in a manual or in research literature indicating that the instrument has acceptable validity and reliability.

- *Interpretability.* Scores generated by the instrument should be easy to interpret for their underlying affective dimensions and should not require complicated or awkward interpretations.

- *Age Appropriateness.* The instrument should be valid for some or all of the ages of students in the study.

- *Administrative Ease.* Classroom teachers should be able to administer the instrument after limited training.

- *Scoring Ease.* Because of the large number of students to be measured, the instrument should be designed for, or lend itself to, machine scoring procedures.

- *Brief Testing Time.* Because the amount of time available for measuring affective behavior is limited, an instrument should not require extensive time to administer.

- *Minimal Response Bias.* Younger students often demonstrate a bias toward socially desirable responses, and instruments should be designed to minimize this bias. This can be accomplished by both the manner in which items are prepared and the types of responses called for.

- *Commonality Across Grades.* Because of the longitudinal nature of the study, either the same or highly related instruments should be used. In particular, instruments prepared in parallel to encompass different grade levels were preferred.

From an original set of over 60 instruments, 12 were judged to be of sufficient merit to warrant further consideration, with advice from a panel of experts. These instruments are described here briefly.

Animal Crackers (Adkins and Ballif, 1975). In its developmental form, known as Gumpgookies. Measures achievement motivation. Developed for pre-school and primary-grade students. Requires 30-45 minutes testing time.

Attitude Toward Learning (Roshal et al., 1971). Measures general attitudes toward learning at school. Developed at upper elementary level. Complicated response format. Requires about 25 minutes.

Attitude Toward Reading (ETS, 1975). Measures attitudes toward reading instruction and reading-related activities. Forms for primary level and upper grade level. May induce positive response bias. Requires 20-30 minutes.

Attitude Toward School (Roshal et al., 1971). Measures attitudes toward school in general. Developed at upper elementary level. Complicated response format. Requires about 25 minutes.

Children's Locus of Control Scale (Bialer, 1961). Measures generalized locus of control. Developed at all elementary-grade levels. Doubtful validity. Requires about 15 minutes.

Children's Self-Concept Scale (Piers-Harris, 1969). Measures concerns children have about themselves. Developed for upper-grade students. Requires 15-20 minutes.

Intellectual Achievement Responsibility Questionnaire (Crandall et al., 1965). Measures control over, and responsibility for, intellectual-academic success and failure. Developed for upper-grade students; has been used at primary grades in national studies. Requires 15-20 minutes.

Locus of Control Scale for Children (Nowicki and Strickland, 1973). Measures generalized expectancies for internal versus external control of reinforcement. Developed for middle and upper grades. Requires about 15 minutes.

M-Scale (Williams, 1972). Measures achievement motivation. Developed for upper-grade students. Doubtful validity. Requires about 10 minutes.

Quality of School Life (Epstein and McPartland, 1975). Measures satisfaction with school, commitment to classwork, reactions to teachers. Developed for upper-grade students. Requires about 20 minutes for full scale.

Self-Concept of Ability (Brookover, 1967). Measures academic self-concept. Developed at secondary-school level, modified for use at primary level. Requires about 20 minutes.

Self-Esteem Inventory (Coopersmith, 1967). Measures attitudes toward self in several domains. Developed for upper-grade students, has been used in primary grades in national studies. Requires about 10 minutes.

The 12 instruments that survived the original screening are shown in Table 12-1, along with an indication of how each instrument fared with the eight criteria described earlier. An 'X' indicates that the instrument was judged adequate on that criterion, '?' means that there was some doubt about the instrument for that criterion, while a blank indicates serious concern.

Recommendations from Panel of Experts. A panel of experts on measuring affective behavior was convened to make recommendations. The panel consisted of: Dr. Joyce Epstein, Center for Social Organization of Schools, Johns Hopkins University; Dr. John Kitsuse, Department of Sociology, University of California, Santa Cruz; Dr. Melvin Seeman, Department of Sociology, University of California, Los Angeles; and Dr. James Vasquez, Far West Laboratory for Educational Research and Development. During the two-day meeting, the panel discussed issues related to the measurement of affective behavior in the SES, examined instruments designed to assess affective behavior, and developed a set of recommendations. The major recommendations were:

1. Use available instruments rather than develop new ones. The panel felt that while

Table 12-1

Evaluations of 12 Affective Instruments

Instruments	Criteria							
	Validity and Reliability	Interpretability	Age Appropriateness	Administrative Ease	Scoring Ease	Brief Testing Time	Minimal Response Bias	Commonality Across Grades
Animal Crackers	X	X	X				X	
Attitude Toward Learning	X	X	X	X	X		X	
Attitude Toward Reading	?	X	X	X	X	X		X
Attitude Toward School	X	X	X	X	X		X	
Children's Locus of Control Scale	?	X	X	X	X	X	X	?
Children's Self-Concept Scale	X	X	X	X	X	X	X	
Intellectual Achievement Responsibility Questionnaire	X	X	X	X	X	X	X	?
Locus of Control Scale for Children	X	X	X	X	X	X	X	
M-Scale	?	X	X	X	X	X	X	
Quality of School Life	X	X	X	X	X	X	X	?
Self-Concept of Ability	X	X	X	X	X	X	X	?
Self Esteem Inventory	X	X	X	X	X	X	X	?

Note: 'X' indicates the instrument was judged adequate on the criterion, '?' indicates there was some doubt, and a blank indicates serious concern.

existing instruments all suffered from some shortcomings, they were undoubtedly superior to any that could be developed in the brief time prior to data collection. Specifically, the panel recommended three instruments: the *Self-Esteem Inventory*, *Intellectual Achievement Responsibility Questionnaire*, and the satisfaction-with-school portion of the *Quality of School Life* inventory. In specifying these three instruments, the panel recognized that they were choosing the best instruments among a less-than-ideal lot, and had particular concerns for validity and appropriateness for younger students. The panel suggested potential modificaions that would be likely to improve each instrument.

2 Read all instruments to all students, in recognition of the obvious fact that the intent in measuring affective behavior is to assess those dimensions without contamination by reading ability While most instruments are intended to be read by the respondent if the respondent has completed third grade, the panel felt that many upper-grade students would lack sufficient skills to handle the reading tasks required; this can be overcome by having instruments read to all students regardless of grade level.

3 Measure affective behavior before measuring cognitive behavior. In considering the total testing schedule for students in the SES, the panel strongly urged that affective instruments be administered prior to cognitive instruments. The panel felt that students' attitudes would be strongly influenced by immediate occurrences, and that if the affective instruments were administered late in the week of testing, students might express negative views that would not be truly characteristic of them but rather would be in response to the (potentially frustrating) achievement-test experiences.

4 Measure students' senses of change over time. The panel suggested that a valuable addition to the assessment of affective behavior would be the use of items that asked the extent to which the student was aware of improvement in skills and changes in feelings and attitudes. In proposing this addition, the panel noted that the longitudinal nature of the study included the repeated measurement of students' affective behavior and thus allowed for the actual consideration of change over time; however, the panel suggested that a particularly useful piece of information would be the student's own awareness of that change. The panel suggested the development of additional items that would indicate the extent to which a student recognized that skills in reading and math had improved, and that feelings about self in the school setting had altered with the passage of time and the experiencing of certain educational activities.

Survey of School Attitudes. The most critical decision, based on the panel's concern with the adequacy of the recommended instruments, was that the study should focus primarily on the measurement of attitudes toward math and reading. Inasmuch as the principal objective of CE programs is to improve skills in reading and math, it was judged most appropriate for the SES to be especially concerned with students' attitudes toward those two curricular areas. Given this decision, a search was instituted to locate the best existing instrument for attitudes toward reading and math.

The search resulted in the selection of the *Survey of School Attitudes* (Harcourt Brace Jovanovich, 1975) for use in the study. This instrument had just recently been developed and released, and is ideally suited to the needs of the SES. The *Survey of School Attitudes* (SSA) is designed to measure student reactions to reading and other language arts, math, science, and social studies. Students indicate whether they like, dislike, or are neutral to different activities in each curricular area. The SSA can be used in group administration settings by a classroom teacher. There are two levels: Primary (grades 1-3) and Intermediate (grades 4-8). The two scales of interest in the SES are:

- *Reading and other language arts:* reading, working with words and sounds, writing, speaking, listening.

243

- *Mathematics:* concepts (of numeration, sets, etc.), computation, geometry and measurement, problem solving, charts and graphs.

The SSA was standardized in 1973 on a sample of 13,500 students in grades 1 through 8. Twelve school systems in ten states participated in the standardization. The standardization sample was highly similar to the nation's population, as indicated by 1970 Census data, on the following dimensions: geographic region, socioeconomic variables, minority population, and community size.

Item analyses of the math and reading scales, based upon combined grade samples, are reported in Tables 12-2 and 12-3.

These data demonstrate that items correlate higher with their own scale than with the other scale, and that there is some degree of commonality in the measurement of attitudes toward reading and toward math.

Table 12-2

Median Item-Scale Correlations From the Standardization Sample

Test Form	Reading	Math
Form A, Reading	54	.37
Form A, Math	.42	.57
Form B, Reading	.56	.38
Form B, Math	.41	.57

Table 12-3

Interscale (Reading and Math) Correlations From the Standardization Sample

Level	Form A	Form B
Primary	.69	.69
Intermediate	.43	.44

Reliabilities of the instrument were determined both by coefficient α and by test-retest with alternate forms over a ten-day interval. Results are summarized in Tables 12-4 and 12-5.

These reliability estimates indicate that the SSA yields reasonably consistent and stable scores.

Demonstration of the validity of the instrument was approached from several directions. To find out whether the instrument actually measures student achievement, correlations

Table 12-4

Internal-Consistency Reliability Coefficients From the Standardization Sample

Test Form	α Coefficient
Primary Reading, Form A	81
Primary Reading, Form B	.83
Primary Math, Form A	85
Primary Math, Form B	.85
Intermediate Reading, Form A	.84
Intermediate Reading, Form B	.82
Intermediate Math, Form A	92
Intermediate Math, Form B	.90

Table 12-5

Test-Retest Reliability Coefficients From the Standardization Sample

Scale	Test-Retest Coefficients
Primary Reading	.65
Primary Math	.65
Intermediate Reading	77
Intermediate Math	.83

between SSA scores and achievement test scores were computed. They seldom exceeded .30, leading to the conclusion that the SSA measures something different from achievement, and is not overly influenced by achievement. To find out whether students respond in socially-desirable directions (to please the teacher), sub-scale intercorrelations were inspected to see if they exceeded scale reliabilities (they did not), and teachers were asked whether they felt students responded honestly (they thought they did). These results do not support the hypothesis of strong response bias. Finally, construct validity was approached through a series of factor analyses, which showed the SSA to measure separate attitude dimensions.

To consider minority-group concerns, a substudy was conducted in which minority and non-minority students' scores were compared. The two groups produced comparable results for both reading and math scales, and on both primary and intermediate levels.

Each item of the SSA contains a picture of an activity related to a curricular area, and has an accompanying statement describing the picture. At the Primary level the statements are to be read by the examiner, at the Intermediate level the statements are also printed on the item The response options are three faces, one smiling, one neutral, and one frowning. By marking the smiling face, the student indicates liking for the activity, marking the frowning faces indicates dislike for the activity, and the neutral face means the student is not sure how he or she feels about the activity.

Modifications to the SSA. Some modifications of the instrument were made for this study Since science and social studies are not relevant, those scales were eliminated. Scales for reading and math were extended by combining items from the alternate forms of the SSA, resulting in two scales, reading and math, with 20 items each. The increased test length was deemed necessary to assess changes that take place longitudinally. The same instrument will be used for both pretest and posttest. This extension of the scales by combining forms is mentioned in the SSA Manual:

> Users who require greater reliability for special applications of the Survey might consider administering both forms. . . . Using both forms would increase reliabilities in accordance with the Spearman-Brown formula, e.g., approximately from .80 to .90 or from .90 to .95.

In addition, a few new items were added to the instrument to create scales that reflect the following attitudes toward school in general, self-concepts in the school setting, and students' sense of change in their own affective behaviors over time. These additional items are described in detail later. The new items, along with the extension of the reading and math subscales, make the length of the total instrument about the same as the original SSA with all four subscales.

We decided that the affective instrument used in the SES would have two levels: a Primary version for grades 1 through 3 and an Intermediate version for grades 4 through 6. Each version would contain 56 items, as follows:

	Number of Items	
	Primary	Intermediate
Reading/Language Arts		
Reading	7	5
Working With Words	6	7
Writing	1	2
Speaking	1	2
Listening	2	1
Other Related Activities	3	3
Math		
Concepts of Numeration, Sets, etc.	9	9
Computation	6	4
Geometry and Measurement	4	4
Problem Solving	1	2
Charts and Graphs	0	1
School-in-General	4	4
Self-Concept in School	4	4
Change Over Time		
School-in-General	2	2
Self-Concept	2	2
Mathematics	2	2
Reading	2	2
TOTAL	56	56

The 20 items for the scale on attitude toward math and the 20 items for the reading attitude scale were obtained by using all 15 items from Form A and five items from Form B. Items chosen from Form B were those that bore the least resemblance to items in Form A, to minimize the extent to which students will feel that they have already answered an item. When assembled into a single instrument the reading and math items were alternated, somewhat reducing the development of response sets by students.

To measure attitude toward school-in-general, four items from the satisfaction-with-school scale of the 'Quality of School Life' inventory were used. The four items reflect attitudes that are not tied to subject-matter areas, adults in school, or other students, but rather deal with school and class in a general manner. The items were modified to allow them to fit a response pattern of 'Yes,' 'Not Sure,' and 'No,' using the same three faces that appear in the reading and math scales. Also, items cast in the negative were changed to the positive to eliminate difficulties students have in responding to negative items. The items are:

I enjoy the work I do in class.

School work is very interesting.

I like school very much.

I am happy when I am at school.

Self-concept in school was assessed by four items selected from existing self-concept scales ('Children's Self-Concept Scale' and 'Self-Concept Inventory') and modified to use the same response pattern. The particular items chosen refer specifically to the student in the school setting, rather than being related to self-concept in broader contexts such as home or play. The items are:

I can think up answers to questions.

I like to learn about new things.

I am good in my school work.

I can learn things quickly.

The eight items measuring the student's sense of change over time in affective behavior were prepared following guidelines suggested by the panel. They, too, were written to use the response mode described above. The items are:

I like reading more than I used to.

I do better work in reading than I used to.

I like mathematics more than I used to.

I do better work in mathematics than I used to.

School work is more interesting than it used to be.

I like school more than I used to.

I am a better student than I used to be.

I like my work more than I used to.

A careful examination of existing locus-of-control scales failed to uncover a reasonable set of items that could be included in the affective instrument. The better scales (e.g., *Intellectual Achievement Responsibility Questionnaire*) are cast in the form of an outcome for which the respondent is to choose one of two possible causes. For example: "If something is easy to learn at school, it is because (a) you pay attention, (b) the teacher gives you lots of help." Using items of this type would require a change in the instrument's response format, adding to the time needed for administration. In addition, younger students may experience difficulty in responding accurately to such items. On the other hand, locus-of-control items cast in the 'yes-no' format (e.g., "When I do good work in class it is because I am lucky") are not the best indicators of internal-external locus of control. Locus-of-control items were, therefore, not included in the new instrument.

PRETEST AND FIELD TEST

When the new items described above were assembled into the second part of the affective instrument, they were randomized to reduce the development of a response set. To ascertain the extent to which students of different ages have difficulties in understanding the new items or in dealing with a change in response pattern, and to estimate the time required for administration, a nine-student clinical pretest was conducted. The new items were presented after five items from the attitude-to-reading scale and five items from the attitude-to-math scale, to provide a realistic trial of the new items. Results indicate that the items are understandable, and that students have little difficulty shifting responses. Time estimates were approximately 30 seconds per item, so that the full 56-item instrument could be administered in less than 40 minutes.

In March 1976, the complete affective instruments were tired out in a field test under conditions similar to those that were anticipated in the formal study. During the field test the suggestions of the panel were followed. the affective instrument was administered on the first day of testing and examiners read the affective items to all students. The two levels of the Student Affective Measure (SAM) were tested in 15 elementary schools that represented a wide range of regional and economic differences. The tests were administered as they would be in the operational years of the study.

The Field-Test Plan

Before the SAM was taken into the field for testing, a detailed plan was developed to guide the systematic analysis of the data so that each critical question would be adequately addressed. The critical field-test questions were concerned with the issues of instrument revision and instrument validation.

Instrument Revision. A number of analyses were planned in the anticipation that the instrument might need revision. The specific questions are listed below:

1 Are the items too complex, conceptually or linguistically, for elementary-school children? Is the testing procedure too difficult? These questions were to be answered on the basis of classroom observations of the field testing by trained observers. A formal observation schedule was developed and provided to the observers as a guide

to use in making observations. If the observations indicated that the students had difficulties understanding the items or the responding procedures, steps would be taken to reduce those problems through revision of the instrument.

2. Do the items have satisfactory response dispersions? The amount of response bias (skewed response dispersion over the range of response alternatives) was expected to be large, but there should be some dispersion of responses over the three alternatives for each item. Any items with no dispersion or too great a response bias across grade levels would be tagged for inspection and possible revision.

Instrument Validation. Two other major concerns for the SAM were the empirical independence of the logical constructs (sub-scales) that are embedded in the measure, and the reliability of the measure and its potential sub-scale components.

1. Should scale scores or one total score be used in the SES analyses? In essence, this question asks whether the logical constructs (sub-scales), described above in the section on modifications, have an empirical independence sufficient to warrant their separate consideration in analyses. The various components of the measure were incorporated because each was thought to be important in the assessment of affective growth. For this reason, it was most desirable that each of the components be considered for independent consideration in the analyses. On the other hand, if there were no empirical support for the independence of the logical components, then separate analyses of them would result in the intractable problem of multiple collinearity of our dependent measures. Two analytic approaches were employed to shed light on this question:

   a. If each of the new logical components (sub-scales) of the measure is separately scored, and the intercorrelations between the components and the total score, corrected for unreliability of each component, exceeds or nearly equals the reliability of the component taken separately, only the total score will be used in the study. This approach asks that the components have greater reliability than total scale coherence in order for them to be considered separately for analytic purposes.

   b. If empirical factors can be formed that loosely permit the conclusion that the components can be 'confirmed,' then the components will be considered to have sufficent empirical support to justify their being analyzed separately for the study. If no such permissive factor structure can be made to emerge, then such evidence will suppport the use of only a total score for the affective measures. (This approach also allows for the weeding out of items that might not cohere in any of the possible empirical factors.)

2. How reliable are the total scale and the sub-scale scores? The reliability estimates, in addition to playing a role in the first validation question, will indicate whether or not additional items will be needed to obtain an instrument of sufficient reliability to be employed in the SES.

The Field Test

The SAM was tested at 15 schools across the country that were selected according to four stratification types: (1) Northwestern, poor, urban; (2) Southeastern, poor, non-urban; (3) Northwestern, non-poor, urban; and (4) Southwestern, non-poor, non-urban. Two classrooms each at grades 2, 4, and 6 at each of the schools were administered the test under standard field-test conditions. A total of 724 second-grade students, 743 fourth-grade students, and 753 sixth-grade students completed the instrument in a valid manner.

Analysis of the Field-Test Data. The data collected from the field test were analyzed by computer and by manual tablulation and content-analysis of information not amenable to computerized analysis procedures. The analyses are presented below, in the same order as the goals and plans for the field test.

1 Are the items too complex, conceptually and linguistically, for elementary school children? Is the testing procedure too difficult?

   - Some evidence of children responding in a uniformly positive manner was observed, but it was not widespread or in response to confusion.
   - Especially at the higher grade levels, test administrators must pace the items more rapidly in order to keep the students' attention.
   - The administration time ranged from 30 to 52 minutes for the second-grade students, from 16 to 34 minutes for the fourth-grade students, and from 17 to 30 minutes for the sixth-grade students.
   - No general problems were observed with the directions to the test.
   - Most items appeared to be understood, even at the second-grade level.
   - The slightly differing response formats in the field-test version caused some problems, but they will be remedied by one consistent response format and item format.

2. Do the items have satisfactory response dispersions? The response frequencies reported in Table. 12-6 indicate that it is not unreasonable to assume that there is a considerable response bias causing the children to mark the happy face, thus earning more positive attitude scores.

   The possibility of response bias in the affective measure was not unexpected. The average dispersions, however, indicated that this bias might not be damaging to the distribution of students over a reasonable range of attitudes.

3. Should scale scores or one total score be used in the SES analyses? The first approach to answer this question was through factor analysis of the items at each grade level in

25A

Table 12-6

Average Response Dispersions for Three Grades

Stimulus	Percentages of Responses		
	Grade 2 (N = 724)	Grade 4 (N = 743)	Grade 6 (N = 753)
Happy-face responses	67 33	62.82	55.79
Neutral-face responses	20.56	24.84	29.24
Sad-face responses	12.98	12.34	14.66
Omitted responses	0.87	0.00	0.31

an effort to see if the separate attitudes presumed to be measured by the hypothesized sub-scales could be said to have some sort of factorial existence. At each grade level, response data for the 56 items were submitted to a principal factor analysis. Item commonalities were initially estimated by the multiple R^2 and then iterated to eight-factor convergence. In order to test the reasonableness of the factor structure in a permissive manner, eight factors were extracted and rotated from each grade level correlation matrix, regardless of other extraction criteria. (If a factor-extraction cutoff had been established at an eigenvalue of 1.00, 13 factors would have been retained for grade 2, 12 for grade 4, and 11 for grade 6.) Targeted rotations to the eight hypothesized factors did not yield acceptably clear factor structures, nor were the factor scales very reliable or independent from one another.

In order to obtain a clear factor solution with more reliable and independent factor scales, a four-factor solution was attempted on the data. The hypotheses were that the four 'change in attitude' factors would be collapsed successfully upon their static analogs, yeilding the following four factors: Attitude to Reading, Attitude to Math, Attitude to School-in-General, and Self-Concept in the School Setting.

The first four extracted factors for each grade level were then rotated to a target matrix composed of the four hypothesized factors. The factors emerged with considerable clarity, but the internal consistency of the last two factors was unacceptably low. Because of the low reliability, the last two factors were assumed to be collapsible, and three-factor solutions were attempted factor-analytically. The clarity of the three-factor solutions depended to a considerable extent upon a 'catch-all' bipolar fourth factor that served as a residual, but the internal-consistency reliability of each of the three factors thus obtained was acceptable. The targeted factor-analytic solutions for three factors and residual for the three grade levels (based on 724 second-grade students, 743 fourth-grade students, and 753 sixth-grade students) are presented in Table 12-7, with hypothesized and targeted factor loadings presented in italics. The three factors were named:

Table 12-7

Four-Factor Targeted Rotational Factor-Analytic Solutions for the Student Affective Measures at Three Grades

Item	Attitude to Reading			Attitude to Math			Attitude to School			Residual		
	Gr.2	Gr.4	Gr.6	Gr.2	Gr.4	Gr.6	Gr.2	Gr.4	Gr.6	Gr.2	Gr.4	Gr.6
1	.36	.34	.42	09	13	11	01	10	18	10	01	- 08
2	31	17	25	.36	.42	.53	13	11	24	06	14	10
3	.40	.42	.40	10	07	12	00	21	17	03	- 10	- 11
4	06	02	22	.47	.39	50	32	19	15	- 06	11	13
5	.48	.42	.43	01	27	25	12	12	10	- 07	- 04	16
6	19	23	27	.44	.32	.49	09	15	14	13	- 02	01
7	.42	.34	36	24	- 01	- 07	15	- 22	30	08	- 10	- 35
8	37	- 08	00	.41	.58	.65	03	31	25	12	03	02
9	.51	.42	.36	04	33	33	06	13	16	17	01	00
10	05	14	19	57	.56	.59	25	18	13	- 02	10	03
11	.26	.54	.56	26	05	18	24	02	- 06	- 09	13	18
12	23	01	01	.43	.54	.57	- 01	24	26	09	07	- 07
13	.47	.46	.45	19	15	12	13	08	09	08	- 08	- 07
14	18	28	23	.44	.34	37	09	05	14	- 01	- 08	- 09
15	.38	33	.37	19	30	24	16	16	30	- 01	- 04	01
16	30	- 02	13	.57	.52	.58	11	28	23	17	02	- 01
17	.51	.41	.44	28	18	32	04	16	14	12	10	10
18	35	21	11	.52	.54	.53	03	17	18	09	09	00
19	.37	32	.33	35	02	01	24	10	22	- 06	00	- 15
20	27	21	22	51	.57	.59	12	15	15	08	11	06
21	.33	.61	.57	24	06	19	07	05	- 02	- 07	16	15
22	17	15	16	.44	.56	.64	12	16	20	- 01	02	03
23	.42	.40	.49	25	15	05	05	14	10	15	06	- 08
24	20	09	10	.58	.66	.70	21	30	21	- 11	00	02
25	.34	.30	.44	39	04	- 04	15	17	26	08	- 05	- 20
26	16	- 32	.43	.51	.47	.52	13	09	14	04	- 02	- 10
27	.36	.33	.41	16	34	21	24	30	29	- 08	01	06
28	20	31	21	.53	.50	.50	10	06	16	06	- 03	- 12
29	.27	49	.49	21	20	29	21	19	14	- 16	18	12
30	14	23	21	.40	.53	.54	15	08	16	03	00	03
31	.44	.34	.27	05	10	07	22	17	16	- 18	- 11	- 27
32	.08	14	20	.50	.60	.70	18	19	24	- 07	10	13
33	.40	.51	.50	29	28	34	20	16	17	11	09	05
34	04	22	14	.62	.55	.54	29	11	20	- 14	- 08	- 12
35	.32	.27	.32	40	- 02	10	21	07	13	18	- 05	- 26
36	21	17	24	.50	.55	.63	05	25	21	05	20	06
37	.46	.43	.43	13	46	38	14	22	19	08	07	12
38	11	16	20	.65	.60	.59	27	21	19	- 07	- 03	04
39	.28	.50	.41	31	16	36	28	24	18	- 03	12	04
40	20	21	19	.47	.49	.51	13	05	17	- 04	- 12	- 15
41	37	27	29	11	10	- 01	.38	.37	.44	- 06	- 15	- 21
42	21	20	22	28	37	41	.47	.53	.44	25	22	21
43	25	18	12	- 03	08	15	.46	.41	.39	- 17	- 28	- 27
44	13	11	22	22	24	21	.32	35	.26	- 23	- 29	- 26
45	11	- 09	- 09	16	26	32	.47	.40	.30	- 27	- 21	- 15
46	20	26	34	26	33	34	.46	.40	.47	31	28	23
47	20	28	32	23	16	28	.30	.24	.33	12	11	- 04
48	13	14	16	13	18	24	.48	.47	.37	- 27	- 31	- 32
49	21	21	18	04	03	09	.43	.34	.33	- 24	- 18	- 29
50	16	22	27	16	25	22	.60	.60	.64	40	35	34
51	16	18	12	18	18	17	.57	.50	.57	- 24	19	13
52	14	12	05	13	13	17	.42	.47	.37	- 39	- 33	- 33
53	02	- 06	- 11	32	37	49	.54	.54	.40	- 07	05	10
54	14	19	27	14	21	12	.56	.55	.58	31	28	20
55	10	21	19	15	15	15	63	.66	.70	30	30	30
56	12	26	13	20	19	28	.62	.63	.66	17	22	16

Note Factor loadings in bold face were targeted and were employed in the interpretation of their respective factors

Attitude to Reading

Attitude to Math

Attitude to School

As a check upon the reasonableness of the factor-analytic solution, scores were obtained for each of the three independent scales and the scales were then intercorrelated within grade levels over all students. As discussed above, these scale intercorrelations should be sufficently small so that there is reason to believe the scales provide assessments of different student attributes. Table 12-8 provides the scale intercorrelations for each grade, the internal-consistency (α) reliability coefficients, and the scale intercorrelations corrected for the unreliabilities of the scales.

Table 12-8

Factor-Scale Intercorrelations (Above Diagonal), Scale Internal-Consistency Reliability Estimates (Diagonal in Italics), and Scale Intercorrelations Corrected for Unreliability (Below Diagonal) for Three Grade Levels

Factor Scale	Grade	Attitude to Reading	Attitude to Math	Attitude to School
Attitude to Reading	2	*.87*	.67	.59
	4	*.85*	.59	58
	6	*.84*	.58	.57
Attitude to Math	2	.75	*.92*	56
	4	.66	*.94*	.63
	6	.65	*.95*	63
Attitude to School	2	.70	.64	*.82*
	4	.69	.71	*83*
	6	.69	.71	*82*

As expected, the scale intercorrelations range in the .50s and .60s, which, when corrected for unreliabilities, range in the .60s and .70s. When the high reliabilities of the scales (.80s and .90s) are considered, it is concluded that each of the three scales has sufficient independence to warrant its independent consideration, especially when employed as a dependent variable in the study of educational processes and services (but it is recognized that the high inter-relatedness could cause analysis difficulties).

Table 12-9, based on the factor-analysis samples, has been included to provide information bearing upon the effects of the high item endorsement rates on the distribution of scale scores. It is clear that all the scales at all the levels are negatively skewed, but in all cases the mean is more than one standard deviation from the maximum score. It is anticipated that the scales will provide a sufficient score range so that potential improvements in attitudes can be reflected with some sensitivity by increases in scores.

Table 12-9

Maximum Scores, Means, and Standard Deviations of Three Factor-Scale Scores for Three Grade Levels

Factor Scale	Grade	Maximum Score	Mean	Standard Deviation
	2	60	50.95	7.01
Attitude to Reading	4	60	50.82	6.75
	6	60	49.04	6.80
	2	60	49.55	8.59
Attitude to Math	4	60	48.48	8.80
	6	60	46 62	9.47
	2	48	41.93	5.98
Attitude to School	4	48	40 88	6.04
	6	48	39 35	6.14

4. How reliable are the total scale and the sub-scale scores? The internal-consistency reliability coefficient for each of the scores is reported in Table 12-8. Total-score reliabilities are .94 at all grade levels.

Conclusions From the Field Test. The major conclusion drawn from the field test and analyses was to retain the affective measure in a form very close to its field-test form. Three scales can be scored and analyzed from the measure: Attitude to Reading, Attitude to Math, and Attitude to School, the last scale being comprised of all the items newly created for the SES.

255

CHAPTER 13. THE EFFECTS OF OUT-OF-LEVEL SURVEY TESTING

Ralph Hoepfner and Ming-mei Wang

Two adjacent levels of the Comprehensive Tests of Basic Skills were administered to each student in a nationally representative sample. The data were used to investigate problems associated with out-of-level testing, floor and ceiling effects, and the identification of low achievers with different test levels. The floor and ceiling effects were found to be related to school concentrations of poor and minority students. It is suggested that these school characteristics might be used as a guide to select better test levels for a school, especially in survey testing where previous school-level achievement data are not available and where individually determined test levels would be impractical.

Out-of-level survey testing has frequently been recommended as an appropriate method for assessing the achievement of students who are not functioning at the level of their grade peers. The rationale is that using the grade-appropriate level with such students will produce only zero or chance-level scores and not truly reflect their (low) skills. To remedy this, several test publishers have made it possible to give a test one or more levels easier, but to reference scores to the students' age-level peers. When employed, the method usually involves the administration and interpretation of achievement tests designed and intended for one or more grades lower than that of the children being assessed. The method rests on the assumption of the unidimensional measurement of achievement growth over the grades and ignores or overlooks potential discontinuities in educational objectives over the grades. If out-of-level testing is effective, it provides a solution to the problem of assessment of disadvantaged, minority, and handicapped students who are often not achieving at the level of their national peers.

One can immediately foresee problems resulting from the implementation of out-of-level testing, however. Already mentioned is the fact that out-of-level assessment is very likely not to assess students on the learning objectives of their current instruction, but instead on the learning objectives of instruction of one or more years previous (but, see Plake and Hoover, 1979). An equally important problem stems from the fact that in the norming operation a test level of a standardized achievement test is seldom administered to students whose assigned grade is far away (more than one level removed) from that for which the level is intended. It is therefore necessary to rely on techniques of curve fitting and on scaled scores. The interpretation of the normed scores must rest not only on the assumption of (unidimensional) scalability of the test levels, but also on that of the functioning equivalence of children over a range of educational development.

PREVIOUS STUDIES ON OUT-OF-LEVEL SURVEY TESTING

These problems notwithstanding, out-of-level testing is still employed in many situations; often enough so that several investigators have attempted to gauge the effects of the method. Ayrer and McNamara (1973) used the results of out-of-level testing in grades 4 to 8 to study its effects. The level of test for each student was determined on the basis of previous grade-level test performance on a different test. The authors noted that among those taking out-of-level tests, the percentage of students scoring below the chance level was much lower than what would be obtained had the grade-level test norms been applied to the scores on the out-of-level tests, with the grade-equivalent score as an intermediary. As the percentage of students in a school taking the out-of-level tests increased, average grade-equivalent scores and percentiles dropped off as much as three months and five points, respectively. This decrease was attributed to the increased floor available and the possible incomparability of normed scores among the test levels, at least for some groups in the extreme ends of the norm distribution.

Yoshida (1976) studied the effects of out-of-level testing with mainstream special-education students by having the teachers select the test levels to be administered on the basis of their inspections of the tests and their knowledge of the students. The out-of-level approach resulted in no reduction of test reliability indexes compared to those of the standardization sample, a large proportion of students scoring above the chance level, acceptable ranges of item difficulties, positive point-biserial item-total correlations, and no apparent ceiling effects.

Employing teacher-selected test levels, Barker and Pelavin (1975) reported that of the students tested two or more levels below the publisher's recommended level, 8.6 percent scored at or above the 80th percentile and 4.2 percent scored at or below the 20th percentile They contended that the interpretable range of scores falls between the 20th and 80th percentiles. Presumably with the rationale that tests become unreliable in the extreme ends of the score distribution, they concluded that the fact that 87.2 percent of the students scored in the interpretable range as compared with the expected 60 percent in the norm group, indicated a success for their out-of-level testing in terms of teacher level selections. It is unclear, however, which grade norms were employed to determine the students' percentiles corresponding to obtained raw scores in the selected test levels. It seems unlikely that norms for the students' nominal grades were used. This ambiguity notwithstanding, the authors seem justified in their confidence that many of the 87.2 percent would have achieved scores so low, if they had been administered grade-level tests, as to vitiate their interpretation. They estimated that with the standard they set, over 55 percent of the students would very likely have earned uninterpretably low scores if grade-level tests had been employed.

The extent of out-of-level testing is quite substantial when teachers are allowed to select levels for each individual student based on known achievement levels. Barker and Pelavin report that test levels below the publisher's recommended level were selected by teachers for 64.8 percent of the students in grades 1 through 6. Of all the students, 26.6 percent were in fact administered tests two or more levels below their grade-appropriate level. Yoshida's study (1976) with a group of special-education students reported a disparity as great as ten grades between the student's age-placed grade and the grade recommended for the out-of-level test selected by the teacher. The investigators of both of these studies felt that the teachers' judgments of appropriate test levels for their students were accurate. If nothing else, the use of teacher-selected test levels appears to be effective in alleviating many difficulties associated with the administration of test levels incommensurate with an examinee's performance level, such as teacher's discontent, student's frustration, and above all, score interpretability. Using two samples of low-achieving students only, Long, Schaffran, and Kellogg (1977) administered the grade-level and instructional level of a commonly used reading achievement test at the fall and spring of the school year. Comparing the grade-equivalent scores corresponding to group mean raw scores in the publisher's norm conversion tables, they found that grade-level testing yielded significantly lower grade equivalents than instructional-level testing at grades 2 and 3, but at grade 4 the reverse was found. This finding was attributed to the inadequacy of the cross-level scaling of the test.

It should be noted, however, that many other factors would have contributed to this phenomenon. Converting mean raw scores to grade-equivalent norm scores is meaningful only if the grade equivalents are a linear transformation of the raw scores. Inspection of the norm tables shows that this is clearly not the case, particularly at the ends of the score distribution. The study sample included only those expected to score in the lower end of the grade-level score distribution and this may reduce the grade-level test reliability for the group A third factor could be that grade-equivalent scores are typically extended to a lower range at the low test levels and thus provide a finer scale for the lower raw scores. The change of signs in the discrepancy between the grade and the instructional-level grade equivalents at grade 4 may be explained by the increased floor effect that is not found at grades 2 and 3.

Long, Schaffran, and Kellogg also found that at grades 2 and 3 more students in the sample were classified as low achievers with the grade-level tests and at grade 4 more were classified as low achievers with the instructional-level test. It should be pointed out that the study explored only one side of the classification problem since only low achievers were studied. It is this side of the problem where the difference between grade-level and out-of-level testing results are most likely to be substantial. In addition, the study results indicate that instructional-level testing generally yields greater fall-to-spring grade-equivalent growth than the grade-level testing for the group of low achievers. Despite some methodological problems, these findings, if substantiated, would reinforce the concern for many evaluators of compensatory-education (CE) programs for the problem of selecting tests appropriate and sensitive to the examinees' achievement levels.

Summarizing the effects of out-of-level testing for the evaluation of CE programs, McLaughlin, Gilmartin, and Rossi (1977) concluded that the use of out-of-level testing can appear to have an effect independent of the effects of the program being evaluated. This can happen because some students simply score at the bottom of the norms, whether they take the grade-level or the below-level test. Since the norms usually extend down further on the below-level test, the students' scores are normed lower than they could be normed with the grade-level test. Thus, changing test levels can increase the apparent deficit of a student by a year or more (and it can decease that deficit in the same manner). The authors recommend that each student be tested with a test level that will provide a score in the mid-range. But heterogeneous groups of students would require a test with several articulated levels and flexible administration procedures.

A FURTHER LOOK AT OUT-OF-LEVEL SURVEY TESTING

In order to implement the congressionally mandated evaluation of the sustaining effects of CE and of Title I programs in particular, we undertook a longitudinal study to describe the nature and evaluate the outcomes of compensatory programs in a stratified-random sample of 242 schools across the nation. At the time of the study's design, there was concern with previous studies of a similar nature that the achievement tests used were either at an inappropriate level, thus providing insensitive if not invalid assessment, or that when achievement was measured over test levels the inadequate inter-level articulation obscured the growth indexes. In order to overcome these potential problems, each of the more than 83,000 students in the representative sample in grades 1 through 6 were tested with two adjacent levels of the Comprehensive Tests of Basic Skills (CTBS) (Hoepfner and Christen, this volume) during the first year of the study.

During the fall of the first year of the study, each student was tested at the test level that the publisher recommended as the best level (grade-level), and was also tested with the level below the recommended one (below-level). This testing schedule would later enable the investigators to select the best functioning level of test for each grade at each school for later years of the study, and create a vertical scale of growth based upon fewer assumptions than those of any published scales. In addition, the data so gathered provide the base for the comparison of the effects on scores of grade-level and below-level testing.

The Sample of Schools. From the 242 randomly selected schools, 111 schools were selected so that most had all six grades in the 1 through 6 range and so that each cell of the sampling matrix was represented by at least one school. The sampling matrix had 84 occupied cells with three stratification dimensions: geography, population density, and poverty. As a check on the possible distortion from national representativeness in the final sample selection, the average percentage of poor students and the average percentage of minority students per school were compared to projected estimates from a large stratified sample of 4,750 schools (Hoepfner, Zagorski, and Wellisch, 1977). In the large sample the projected estimates were 28.39 percent poor students per school and 23.79 percent minority students per school. In the reduced sample of 111 schools the average percentages were 27.23 and 19.47, indicating a slight undersampling of schools with higher concentrations of poor and minority students.

For each school, two indexes were computed for each grade and each level of test: the percentage of students scoring at or below the chance level (defined as the number of items divided by the number of alternative responses to each item), and the number of students with perfect or one-less-than-perfect scores. These two indexes can clearly be seen to be operationalizations of 'floor' and 'ceiling' effects for the tests.

The Questions Investigated. With the data for each grade and test level at each of the 111 schools, several questions of interest were investigated. The relationship between the percentage scoring below chance for each level would provide an answer to the question of whether below-level testing reduces floor effects. A similar analysis employing the percentage of perfect and near-perfect scores answers the question of whether below-level testing increases ceiling effects. Both answers, further, could be conditioned on the percentages of minority and poverty students to provide better insights into the effects. The effects could also be investigated separately for reading and math scores, as both were available.

Because the test data had been completely re-normed on a closely representative sample (the 242 schools) and with inter-level scaling problems directly solved through the double testing procedure, the study was also able to provide evidence relative to the assertion by Long et al. (1977) that the percentage of low-achieving students could be manipulated systematically through the selection of grade-level or below-level testing.

STUDY FINDINGS

Floor and Ceiling Effects by Test Level. The mean school percentages of students scoring at or below the chance score, and of students with perfect or near perfect scores, are presented in Table 13-1, by grade and by test level. It is clear that in all cases for both reading and math tests the use of below-level tests reduces the percentages of students scoring at or below the chance score. The reduction is remarkably large at grade 1. Except for grade 5, the grade-level tests result in an average of over 10 percent of the students scoring in the floor range. The ceiling effects are much less severe. In particular, at grades 4, 5, and 6 average percentages of students scoring at or near perfect are generally small (less than 2.5 percent) with only one exception. At grades 1, 2, and 3 the ceiling effects are considerably

Table 13-1

Average School Percentages of Reading and Math Scores Below the Chance-Level Floor and at or Near the Ceiling, by Grade and Test Level

		Reading Scores		Math Scores	
Grade	Level	Average School Percentage Scoring at or Below Chance	Average School Percentage Scoring at or Near Perfect	Average School Percentage Scoring at or Below Chance	Average School Percentage Scoring at or Near Perfect
1	Grade	53.2	1.0	50.4	0.1
	Below	10.5	14.1	2.6	4.9
2	Grade	12.2	3.4	12.3	0.5
	Below	7 0	15.9	7.1	5.2
3	Grade	16.2	0 1	11.8	0.1
	Below	2.3	16.4	2.4	6.3
4	Grade	14.1	0.0	10.4	0.0
	Below	7 0	0.6	2.5	2.3
5	Grade	7 7	0 1	3.1	0.0
	Below	2.6	2.2	0.8	9.3
6	Grade	12.0	0.0	17.4	0 0
	Below	3 9	0.4	1.7	0.0

smaller for math than for reading with below-level tests. There are only negligible ceiling effects for all grade-level tests.

As McLaughlin et al. remark, the floor effect is the more important problem, particularly when testing CE students. In view of the present findings, if school specific recommendations are not available, below-level tests of the CTBS are more likely to be appropriate for reading and math.

Floor and Ceiling Effects by Poverty and Minority Concentrations. The schools' floor and ceiling percentages for each grade and for each level of the reading and math tests were related to concentrations of poor and minority students. Although all the variables are expressed as percentages, their frequency distributions were generally quite skewed. Pearson product-moment correlations were nonetheless employed to show the relationships because attempts to categorize the variables into more analyzable ordinal form all too often obscured the extent of the underlying relationship. The relevant average correlation coefficients are presented in Table 13-2.

Table 13-2

Average Correlations Over Grades of Extent of Floor and Ceiling Effects With School Concentrations of Poverty and Minority Students

	Poverty Concentration		Minority Concentration	
	Reading	Math	Reading	Math
Extent of floor effects with grade-level tests	65	56	.63	.51
Extent of floor effects with below-level tests	67	.49	.62	.44
Extent of ceiling effects with grade level tests*	–.22	– 16	–.15	–.08
Extent of ceiling effects with below-level tests	–.42	– 39	–.34	–.30

*The average correlations for reading are obtained for grades 1 to 5 only, as no schools had any grade 6 students scoring at or near perfect for the grade-level test In fact, there are non-negligible variations across schools in the extent of ceiling effects with grade-level tests only from grades 1 and 2. For grades 3, 4, and 5, the variation is extremely small, with most schools having no students scoring at or near perfect Due to such a lack of variation and the extreme skewness of the distributions, the low correlations are expected and do not necessarily reflect a lack of relationship between the extent of ceiling effects and the concentration of poverty and minority students.

Likewise, the average correlations for math are obtained from grades 1, 2, 3, and 6 only, as the ceiling effect indicator is zero for grades 4 and 5 in all the schools. For all other grades, the variations are very small. Again the distributions are highly skewed for these grades. As a result, no inference about the relation between the extent of ceiling effect with grade level test and the concentrations of poverty and minority students should be made. Nevertheless, the average correlations are presented here for the completeness of the table.

The percentage of students scoring at or below the chance score correlates positively and substantially with school concentrations of poor and minority students. With both grade-level and below-level tests, these correlations are greater for reading than for math. A negative and smaller relationship is observed for the percentage of students scoring at or near perfect with below-level tests and the school's concentration of poor and minority students. This relationship is again slightly stronger in reading than in math. Since there are very small ceiling effects with grade-level tests, the relationships in these cases are of less interest (see footnote to Table 13-2).

In large-scale survey testing, an evaluator cannot ignore the floor effects—resulting from giving inappropriate test levels—on the sensitivity and validity of the data analyses. The results reported here suggest that knowledge of a school's concentration of poverty and minority can be utilized to aid in the selection of test levels to be administered.

Between-Level and Between-Skill Area Correlations. Because there is concern with the problem of floor effects, it is of further interest to investigate the relationships between the floor effects of reading and math tests, and between the floor effects of grade-level and below-level tests. The average correlations between the floor effects of reading and math tests are .74 and .65 for grade-level and below-level tests, respectively. The correlations are comparable to the correlations between reading and math scores on the CTBS. The floor effects of the two levels of tests are also highly correlated, with averages of .74 for reading and .67 for math. The between-level correlations of floor effects are slightly lower than those of the test scores.

Classifications of Student Achievement Status by Test Level. In order to verify the findings of Long et al. (1977), the nationally projected percentages of students who would be classified as low achievers on the basis of scores one year or more below the expected level in terms of grade equivalents were determined separately for grade-level and below-level tests and also separately for reading and math. The percentages are presented in Table 13-3.

For reading, at every grade but the first, more students would be classified as low achievers by the grade-level tests than would be by the below-level test. For math, the same findings hold except that at grade 2, the below-level test appears to classify more students as low achievers. The difference, however, is so small that it is likely due to the discrete nature of the test scores and the rounding to integers. At grade 1, the cutoff for low-achieving scores is approximately half a year, instead of the usual one year below the assigned grades (see footnote to Table 13-3 for explanations). The exception at the first grade could be explained by the severe floor effects of the grade-level tests and the necessity of extrapolating the grade-equivalent scale into the truncated part of the score distribution. This phenomenon greatly reduces the percentage of students classified as low achievers with the grade-level test as compared to that with the below-level test.

The grade-equivalent scale was developed from the vertical scale scores which, in turn, were calibrated by joint utilization of the test results from all grades and both test levels at each grade. Based on the grade-equivalent scale developed in this manner, the scores corresponding to the cutoff for the low-achieving category closely approximate the empirically determined performance levels of the typical students (i.e., median performer) from one

Table 13-3

Projected National Percentages of Students Scoring One Year or More Below Expected Grade-Equivalent Score, by Grade Level, Test Level, and by Reading and Math Scores

Grade	Percentage One Year or More Below Grade Level			
	Reading Total Score		Math Total Score	
	Grade-Level	Below-Level	Grade-Level	Below-Level
1*	7	14	9	13
2	8	6	7	8
3	25	17	16	15
4	31	28	20	17
5	32	31	25	18
6	38	33	33	27

*For grade 1, the percentages are of students scoring approximately one-half year or more below the 1.1 grade equivalent Half a year is adopted because it corresponds to a similar percentile below which grade 2 students will have grade-equivalent scores one or more years below 2.1 in terms of the grade-level norms.

grade lower. Thus, the finding of more low achievers in terms of grade-level tests at grades 3 to 6 cannot be simply attributed to the scaling procedure. Nor can it be accounted for by the greater difficulties of the grade-level tests. Rather, it probably reflects the skewed distributions of the test scores and the differences in the score distribution forms of the two test levels.

The present results, based on the complete range of students, clarify the inconsistencies of the findings at grades 2, 3, and 4 reported by Long et al. (1977). The present data point to the conclusion that grade-level testing results in greater percentages of students considered as low achieving. Since the grade-level tests would be expected to better reflect the content and the curriculum for most students tested, it would appear that grade-level tests can more accurately detect a student's low-achieving status, though they identify more low achievers in the population.

The data also confirm the claim by McLaughlin et al. (1977) that greater achievement deficit in terms of grade-equivalent units can be obtained with lower test levels. This is largely because the grade-equivalent score corresponding to the chance-level score increases with test levels. A pronounced case can be seen at levels 2 and 3 of the CTBS. It was found, based on the new norms, that if a sixth grader obtains the chance-level scores on both test levels in reading, a grade equivalent of 3.4 will be earned with level 3, but only 2.4 will be earned on level 2. Consequently, the student will be judged to be one more year behind the level 2 test. In this sense, grade-equivalent units in general are not ideal measures of achievement deficiency.

CONCLUSIONS

This study reveals that the use of below-level tests reduces the percentages of students scoring below chance levels. It also reduces the percentage of students scoring one year or more below their expected grade levels, with the only noticeable exception at grade 1. The floor effects were found to be greater in schools with higher concentrations of poor and minority students. The ceiling effects, on the other hand, were small for all grade-level tests. With below-level tests, considerable amounts of ceiling effects were found at grades 1 to 3. The relationship of below-level ceiling effects and the poverty and minority concentrations is negative and weaker than that observed for the floor effects.

CHAPTER 14. SPEEDEDNESS OF ACHIEVEMENT TESTS IN THE SUSTAINING EFFECTS STUDY

Ralph Hoepfner

The test of academic achievement used in the Sustaining Effects Study (SES) was examined for differential effects of speededness among racial ethnic groups. Although small differences in speededness were detected, it is not at all clear that the differences result in bias in the measurement of achievement that will later have a detrimental impact on study findings. The reason for this, it is argued, is that there is no compelling argument for the exclusion of speed of thinking and responding in measures of the outcomes of compensatory education.

Because the SES was expected to provide answers to questions of educational policy, it was planned from the beginning that shortcomings of previous evaluations would be avoided. For this reason, a great deal of care went into the selection of the tests of academic achievement (see Hoepfner and Christen, this volume). Sufficient time and effort were allocated to ensure that the achievement measure would result, as much as possible, in clear findings. The study, for example, built its own fall and spring norms, so that indexes of growth would not depend on interpolations or extrapolations. Vertical-scale scores were developed for the study that capitalized on the semi-annual administrations and that did not depend on assumptions of equivalence between different groups of students (Hemenway et al., 1978). Items exhibiting statistical bias were removed from the scores (Hoepfner and Christen, this volume). In other words, many of the traditional assumptions or practices of achievement testing were not uncritically accepted.

In part, this critical approach was based on concerns regarding how the achievement of deprived students can be fairly assessed, because the focus of the study was to be the achievement growth of (what turned out to be) Title I students over more than one school year. Problems of item bias and out-of-level testing were faced and resolved early (Hoepfner and Christen; Hoepfner and Wang, this volume), as there was a history of concern with them. The issue of the effects of speeded tests on the results for deprived students had not enjoyed such historical concern, but the issue was raised early in the study, and it had to be resolved prior to the planning of test administrations.

CONJECTURES ABOUT SPEEDED TESTS

The lack of previous empirical findings in this matter led several advisors to the study to conjecture regarding how the use of speeded tests might bias or obviate the findings. These conjectures can best be understood as arising from issues of bias, validity, and logistics.

Speeded Tests and Biased Measurement. Standardized achievement tests have been developed so that the established time limit permits the typical student to attempt or complete all the items. Most definitions of test speededness assume that some small percentage of examinees will not have time to complete or attempt all items. If this small percentage is non-randomly distributed among certain groups of students, then conditions of speededness introduce bias against members of those groups.

It was suggested that disadvantaged or deprived students may be just such a group. If speeded tests showed greater socioeconomic or even racial/ethnic differences than unspeeded tests do, such information would lend credence to the suggestion. In this event, the study would have to take strong steps to reduce or eliminate the biasing factor so that the conclusions drawn from the study, concerning the effectiveness of programs designed for those deprived students, would not be biased in a manner that reduces the observed growth and leads to a conclusion that the programs are not effective.

The logic behind the supposition that scores from timed tests can contribute to bias is that the rate and accuracy of responding depend on cultural, personality, and motivational factors as well as on ability. If accuracy were equally distributed among racial/ethnic groups, but rate were not, then a timed test would be biased against those groups with slower rates, even if they had equal accuracy. The slower rates would result in lower total scores on timed tests, thus penalizing the slow but accurate responder. Khan (1968), studying high school students, found that time limits influence scores of black students to a greater degree than white students. He concluded that speededness may bring out cultural differences, and may, therefore, reduce the validity of a test. Because Title I students are disproportionally minority students (Wang, Hoepfner, Zagorski, Hemenway, Brown, and Bear, 1978), we might expect speeded tests to distort their actual achievement.

On the other hand, Berry and Lopez (1977), in a comprehensive review of testing problems for Spanish-speaking students, failed to mention speededness as a culturally biasing factor. When Flaugher and Pike (1970) administered tests that were too difficult to inner-city high school students, although scores were very low, measures of speededness were similar to those based on a nationally representative sample. Their findings strongly suggest that difficult achievement tests are not differentially speeded for minority groups. Attempting to reduce expected bias caused by speededness, Evans and Reilly (1972, 1973) found that scores for both black and white college seniors increased, but not differentially. They concluded that reducing speededness in tests was not beneficial to the black examinees insofar as the increased time did not reduce score differences between the minority and majority groups. Using only bright students, Bridgeman (1980) showed that students who are quick at one task may not be quick at others, that there is no single and general trait of quickness that can, at present, be ascribed to any group of individuals, no matter how they are characterized.

Yates (1966) found that some nine-year-old children's scores are more seriously affected by speededness than others, but no personality factors could be found associated with those differences. Likewise, Wasson (1969) found that extending the time limits for a reading test for fifth-grade students resulted in increasing total scores, but the increases were not associated with characteristics such as tested intelligence, or even reading speed. Miller and Weiss (1976), found that examinees of different ability levels show similar patterns of response in adapting to timed or untimed testing conditions.

The findings and conclusions from these earlier studies, while not confirming or denying biasing effects of speeded test conditions, suggest that the SES ought to investigate them further, so that the issue might be resolved and the study thereby improved.

Speededness and Validity. Use of speed-dependent outcome measures could affect both the internal and external validity of a study. To the extent that the degree of speededness varies among test forms and administrations, estimates of actual growth will be influenced. The results are confounded even more if the test scores from the groups being compared are characterized by differential speededness. Either of these occurrences could affect interpretations of the results of the study.

With respect to the effects of speededness on the external validity of the study, we must question the extent to which speeded performance is the goal of compensatory education programs or, more generally, if such performance will help the students in such programs to rise out of their deprived conditions. As long as the instruction is speeded, the learning of students will depend on mental quickness. (This offers a partial explanation of why tests are such valid predictors of academic performance.) However, if speededness is removed from the instruction, then a speeded test will be less valid as an indicator of effectiveness of instruction To the extent, then, that quickness of response is not something that will improve as a result of CE programs, it should not be part of the indicator of its effectiveness. (One could question, further, if mental quickness ought to be at least an implicit effect on the grounds that it may be important for the students in functioning in the non-academic job market they are likely to try to enter, but such questions seem highly speculative.)

Psychologists have not definitively resolved the problem of how speededness should be considered in relation to intellectual performance. We simply have no supportable and general answers to questions regarding the possibility of thinking slowly but well, if slow performance is mainly a reflection of long latency in initiating thought on a problem or slowness in plowing through the stages of a solution, if it is caused by checking and rechecking possible solutions, or if it merely indicates slowness of responding after a solution has been achieved. Horn (1979) found that the correlation between number-correct scores and speediness of response is generally low (about +.22) and he concludes that they are largely independent. His data imply that slowness in providing answers is not necessarily indicative of providing poor answers. Speediness, of course, is a factor that appears to run through all forms of intellectual and performance decline with aging (Birren, 1974).

Speededness and Testing Logistics. The conditions under which a test becomes speeded primarily include administration under rigid time limits. Variations in the time limits will result in differences in mean scores. Since it is very difficult to achieve uniform adherence to time limits, one could consider untimed tests in which each student is allowed as much time as needed to work on the test (see Rindler, 1979, for other approaches to assessing speededness).

This universal solution to the problem of speededness would eliminate the suspicion that some test-wise teachers manipulate testing time to influence their students' scores. Whether or not teachers of deprived children allow them a bit extra time, or whether they reduce time for pretests and increase it for posttests, or whether they reduce it to improve chances to qualify for special compensatory services, is a matter of speculation. Granting each student the time needed would eliminate those confoundings, but would present logistic problems in test administration At the ages where test items are read aloud (generally grade 2 and lower), the pupils could raise their hands when they have responded, so that the students, not the administrator, set the pace But just how patient would administrators be before they made comments designed to hurry the slow students along? One also has to consider the peer pressure of the group that is waiting for the next item, to say nothing of tired arms and wandering attentions.

At the higher grades, it may be difficult to keep the faster students occupied and well-behaved under truly unspeeded conditions. Assigning them additional test items may be seen as an unfair burden on their cooperation. It should be clear, therefore, that the solutions to the potential problems of speededness may not provide a net gain for the SES.

THE SPEEDEDNESS OF THE ACHIEVEMENT TEST FOR THE SES

The technical reports (CTB/McGraw-Hill, 1974) that support the Comprehensive Tests of Basic Skills (CTBS), the test chosen as the measure of student achievement for the study, are not specific about the degree of speededness of the tests or if the speededness is different for any identifiable group of students The publisher did supply us with data tapes selected from its norming sample that provided us the item response scored as right, wrong, or omitted for each test level and grade. The item data for each student was also associated with a racial/ethnic identification of the student into 'black,' 'brown,' and 'other' (predominantly white) By tabulating the incidence of omitted items at the end of the test, we could determine if any of the test levels were speeded. The approach will give somewhat inflated estimates of speededness, however, because students are not encouraged to guess on items, and some of the omitted responses may be due to lack of knowledge and reluctance to guess, instead of insufficient time to address the item.

Speededness, of course, is a relative thing. In order to give it some specific meaning, criteria have been set by which one can classify a test as speeded. These criteria generally use one or more of the following test characteristics: (1) the percentage of examinees attempting the last item of the test, (2) the percentage of examinees completing the first 75 percent of the test items, and (3) the percentage of items attempted by at least 80 percent of the examinees. The tabulations based on a random sample of the publisher's data tape are presented in Table 14-1. The reader should keep in mind, however, as noted above, that the values in Table 14-1 probably over-estimate the speededness of the tests.

The values in Table 14-1 indicate that speededness is greatest for the Math Computation scales. Tests of this kind are frequently speeded, in part because the domain of items is so large and the items so easy to create, and in part, presumably, because speed of calculation does have some practical value The Math Concepts scale is least speeded. Speededness appears greater at grades 4 through 6 than at the earlier grades, but this change does not parallel the change from oral to silent administration of the scales. Although by most standards the data in Table 14-1 would be interpreted as indicating that the CTBS scales are not

Table 14-1

Three Indexes of Speededness of the CTBS Scales, by Grade

Scale* Grade	Percentages Attempting the Last Item	Percentages Completing the First 75% of Items**	Number of Items Attempted by at Least 80% of Examinees
Reading Vocabulary			
Grade 1	96.2	98.9	100.0
Grade 2	96.6	97.2	100.0
Grade 3	99.2	99 9	100.0
Grade 4	86.3	90.2	100.0
Grade 5	86.2	92.5	100.0
Grade 6	82.0	90.2	94.2
Reading Comprehension			
Grade 1	96 4	97 6	100.0
Grade 2	86 1	90 0	100.0
Grade 3	94 5	94.7	100.0
Grade 4	88 1	91 8	99.2
Grade 5	86.8	92 6	100.0
Grade 6	86.9	93.6	100.0
Math Concepts			
Grade 1	90.7	98.2	100.0
Grade 2	96.7	97.5	100.0
Grade 3	94.5	98.0	100.0
Grade 4	95 5	96.6	98.7
Grade 5	97.4	98.4	100.0
Grade 6	96.9	98.3	100.0
Math Computation			
Grade 2	82.5	87.5	100.0
Grade 3	95.6	97.8	100.0
Grade 4	92.7	95 7	100.0
Grade 5	78.2	87.4	96.5
Grade 6	81.5	91.4	95.8
Weighted Totals			
Reading Vocabulary	87.9	93.1	97.6
Reading Comprehension	88.4	93.2	99.9
Math Concepts	96.0	97.9	99.8
Math Computation	84 2	91.4	97.6

*The Reading Vocabulary scales in grades 1 and 2 are named 'Sound Matching' and 'Word Recognition II,' respectively, the Reading Comprehension scales in grades 1 and 3 are 'Letter Sounds' and 'Comprehension Passages.' At grade 1, only one math scale is given, named 'Mathematics,' which we have classified as a Math Concepts scale for this table and for Figure 14-3.

**The percentages reported in this column are the smallest percentages of omitted responses for any item in the last fourth of the scale's items.

speeded, there is some speed which may have more than a negligible impact on study findings if it is distributed in a manner to confound analysis.

DIFFERENTIAL SPEEDEDNESS AMONG RACIAL/ETHNIC GROUPS

The students from the publisher's norming group were not identified by their participation in compensatory programs, so direct tests of speededness could not be made among the comparison groups to be used in the study. There was information on the race/ethnicity of the students on which differences in speededness could be observed, however, so the percentage of responses omitted by students in each racial/ethnic group was tabulated for each item of each scale. In order to make the rates of omitted responses visually comparable, the items in each scale were calibrated so that all scales would appear to have 100 items. Then the plots of the percentages of omitted responses by item were smoothed by the method of moving averages. The results are presented in Figures 14-1 through 14-4, for each of the CTBS scales.

The important things to look for in the figures are curves that rise at different rates for the three groups (if the curves are level or declining, the omissions are not likely due to conditions of speed, but to disinclinations to respond). In grades 1 through 3, we can see confirmation of the data in Table 14-1: speededness is not apparent, but response omissions occur for all the items. At these grades, the 'other' students have lower rates of omitted responses than the black or brown students, but the differences in rates are frequently not large. Except for the Math Concepts scale, speededness becomes apparent in the higher grades. In general, we can conclude that speededness is most pronounced for black students at the higher grades and less pronounced for brown students.

The differences in rates of omitted responses become quite large at times (as much as 15 percent), so differences in the observed scores will be noticeably influenced by them. Because it seems safe to assume that the differences in rates of omissions are largely due to speededness, we conclude that speededness will play a role in the analytic comparisons of the study. This role will be limited to the correspondence between the racial/ethnic groups and the comparison groups of students formed on the basis of their participation in CE programs, but it is difficult to assess exactly because while greater percentages of minority students participate in CE programs, more majority students, in absolute numbers, participate Our conclusions regarding the meaning of the effects of speededness still depend on whether we accept test-taking speed as a reflection of mental speed, and then whether we believe that mental speed is a desirable outcome of compensatory services. To the extent that we do believe it we have increased confidence in the external validity of the measures of academic achievement.

2/0

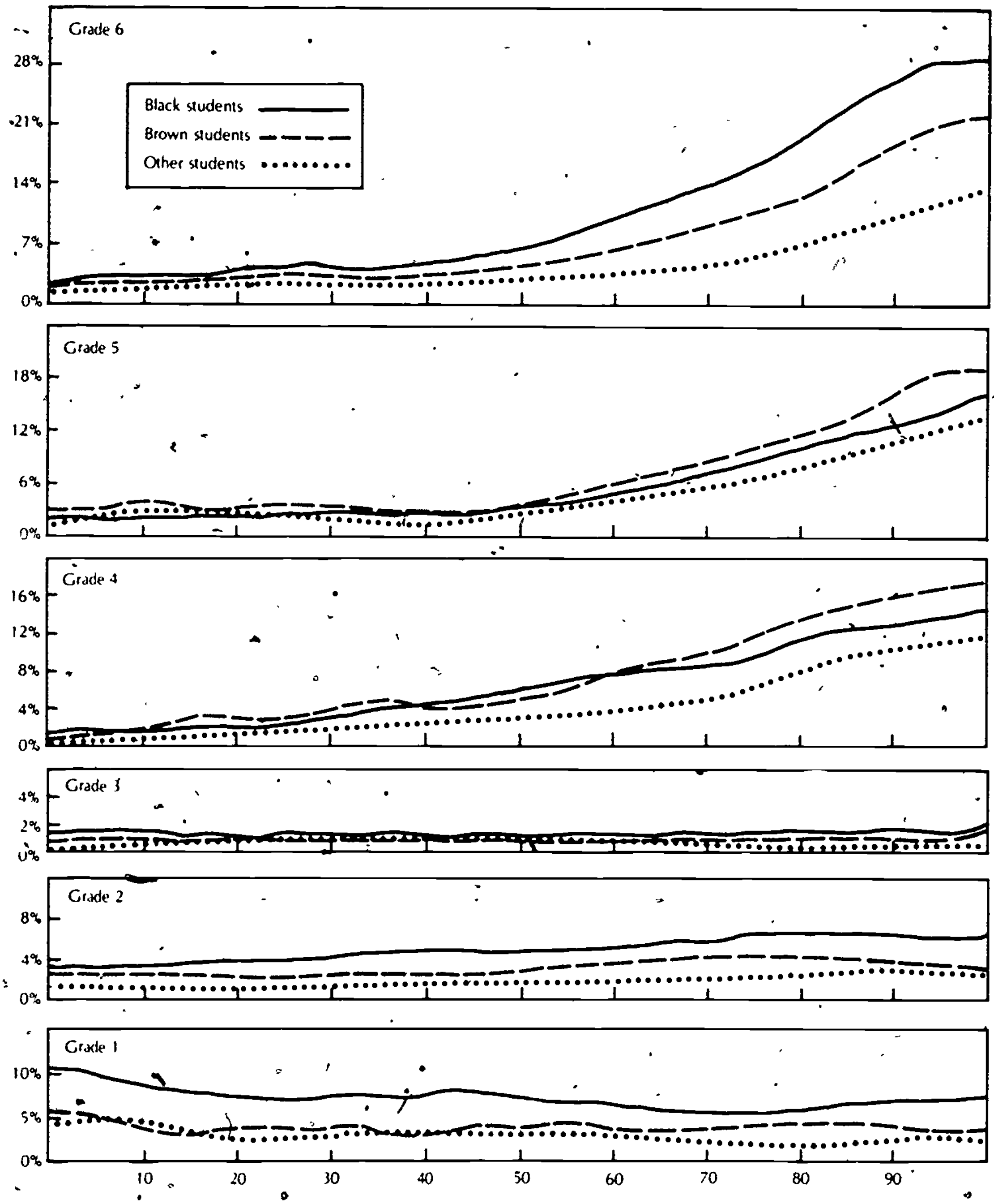

Figure 14-1

Smoothed Percentages of Omitted Responses (Vertical Axes) for Items of the CTBS Reading Vocal[illegible]ary Scale (Horizontal Axes, Scaled so All Test Levels Appear To Have the [illegible]e Number of Items) for Three Racial/Ethnic Groups, [illegible]s Indexes of Speededness.

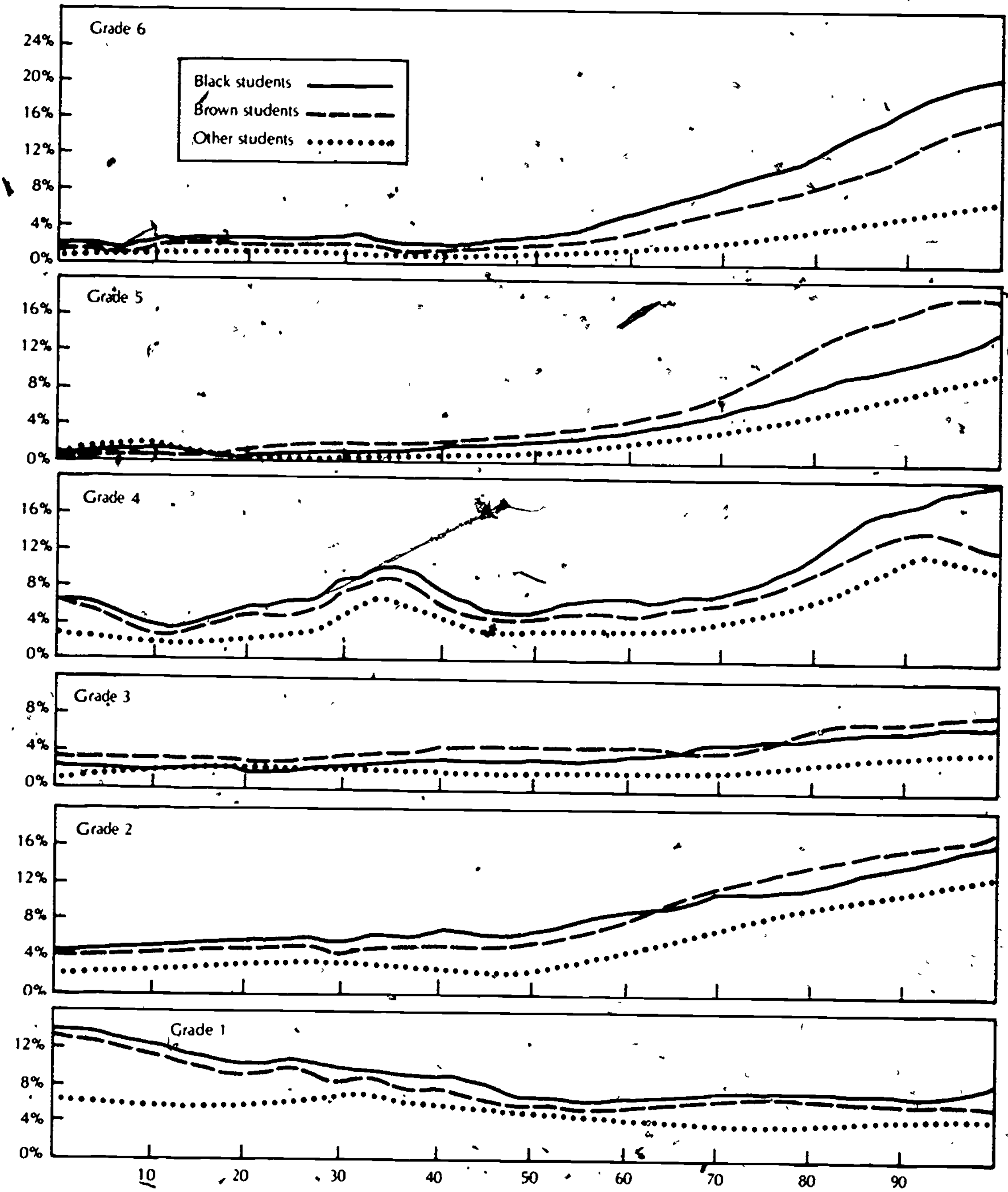

Figure 14-2

Smoothed Percentages of Omitted Responses (Vertical Axes) for Items of the CTBS Reading Comprehension Scale (Horizontal Axes, Scaled so All Test Levels Appear To Have the Same Number of Items) for Three Racial/Ethnic Groups, as Indexes of Speededness.

272

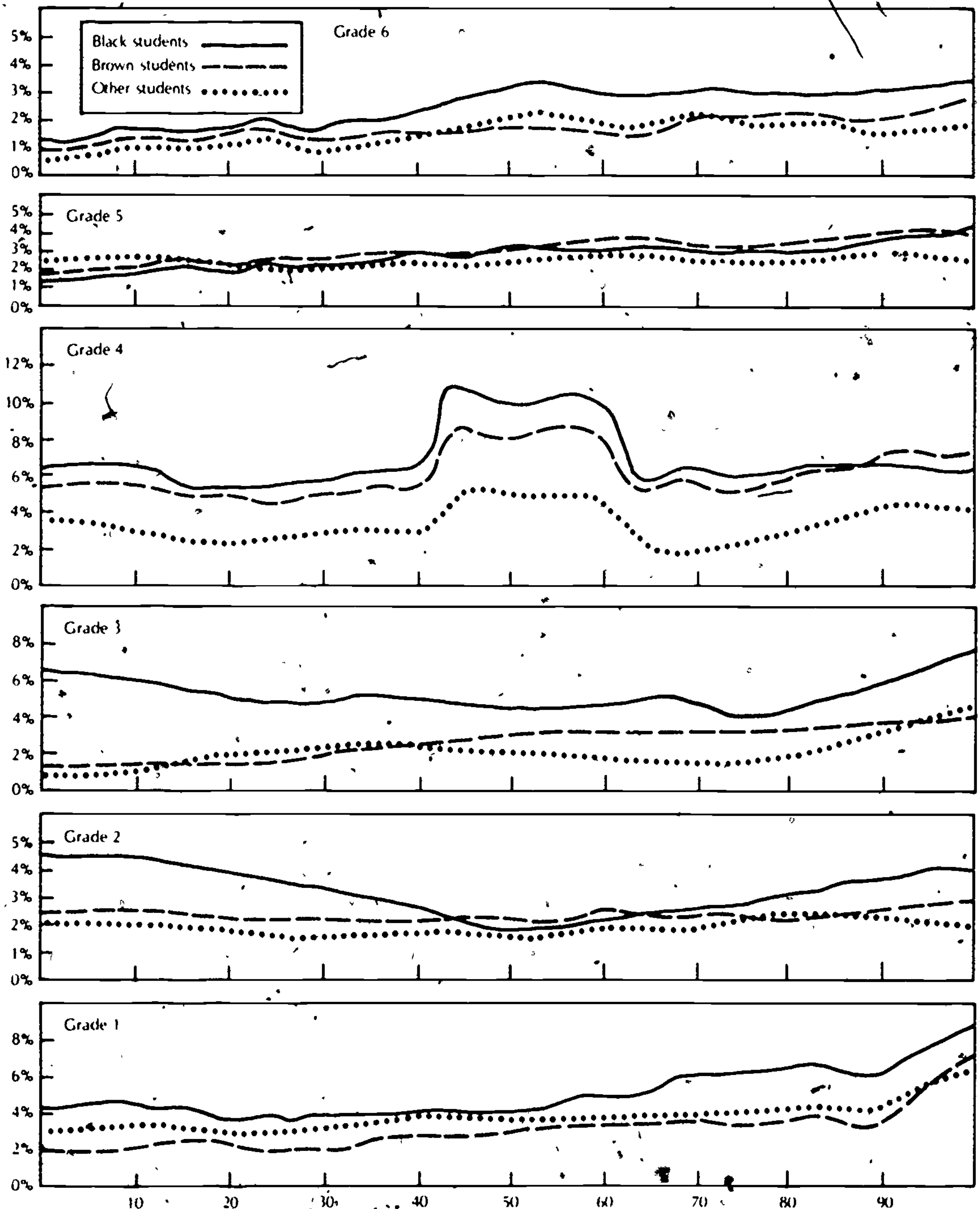

Figure 14-3

Smoothed Percentages of Omitted Responses (Vertical Axes) for Items of the CTBS Math Concepts Scale (Horizontal Axes, Scaled so All Test Levels Appear To Have the Same Number of Items) for Three Racial/Ethnic Groups, as Indexes of Speededness.

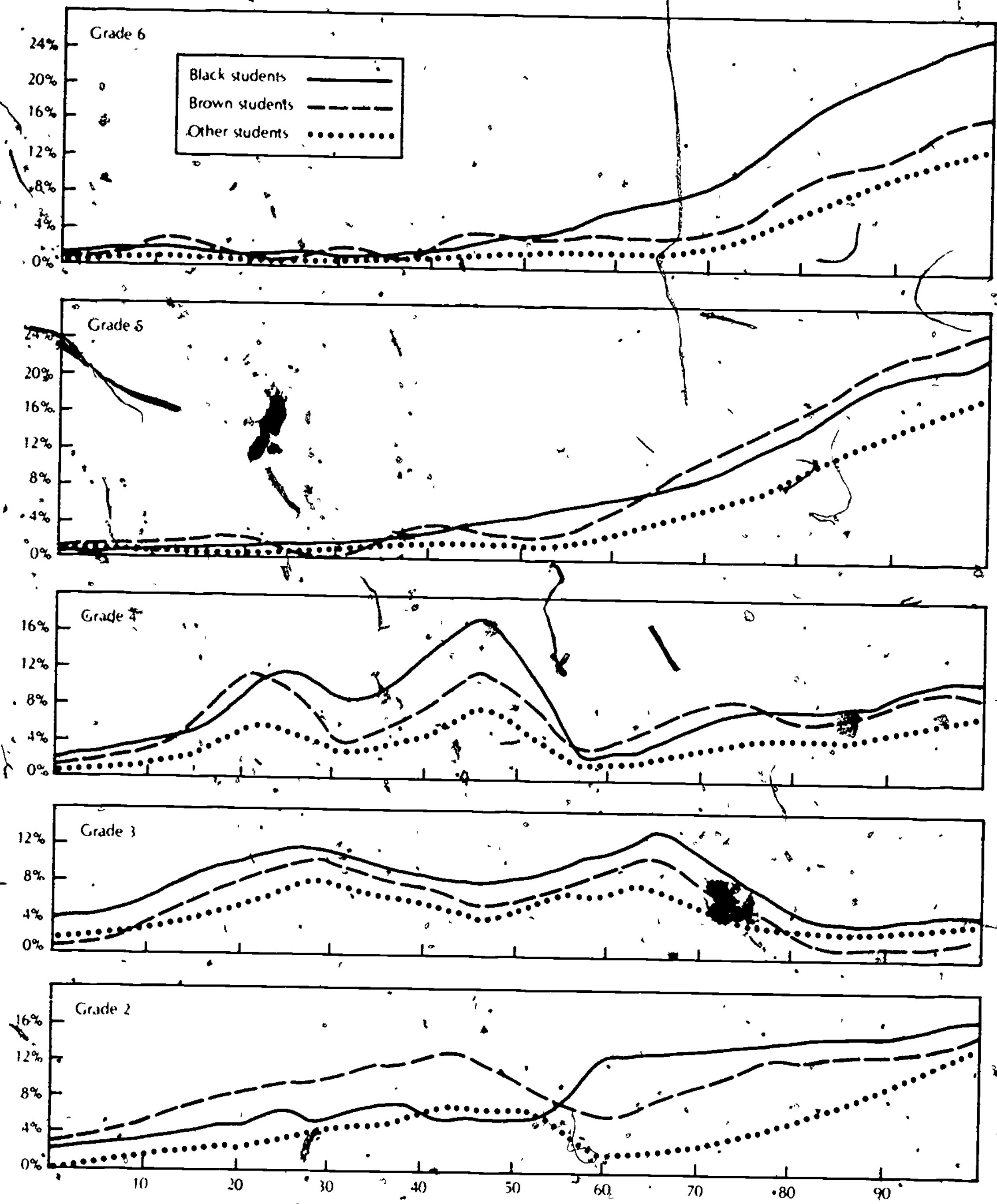

Figure 14-4

Smoothed Percentages of Omitted Responses (Vertical Axes) for Items of the CTBS Math Computation Scale (Horizontal Axes, Scaled So All Test Levels Appear To Have the Same Number of Items) for Three Racial/Ethnic Groups, as Indexes of Speededness.

PART IV. THE EFFECTS OF ATTRITION

Part IV is composed of two chapters that address two issues of attrition of cases from the longitudinal study. In the first chapter, the reduction in the study's sample of schools is described. Although the retained sample was selected with specific purposes in mind, statistical comparisons show its ability to represent the nation very well. Some potential complications for the study analyses resulting from the systematic sample reduction are discussed.

Based on the first full calendar year of the Sustaining Effects Study, we document, in the second chapter, the incidence of attrition and compare rates over several dimensions of student chracteristics, such as minority status and achievement level. We also study the attrition rates by several characteristics jointly, in order to understand better the major sources of attrition. Finally, we provide some conjectures about the expected influences of the observed attrition on different kinds of analyses and issues in the study.

CHAPTER 15. THE REDUCED SAMPLE FOR THE SECOND AND THIRD YEARS OF THE SUSTAINING EFFECTS STUDY

Ralph Hoepfner

The representative sample of schools for the first year of the Sustaining Effects Study (SES) was created large enough to ensure accurate national projections of the incidence of Title I, so that Congressional mandates for highly accurate national estimates could be met. In the second year of the study, when national projections were no longer planned, the sample was reduced. Sample reduction was accomplished in several ways that could have resulted in biases that might influence the analyses of the longitudinal effects of Title I. Therefore, several tests were made of the reduced sample to assess the bias. The results of the tests indicated that, even though the sample reduction had not been random, the reduced sample was representative of the nation's schools and could be expected to support inferential analyses with very little distortion.

The complete first-year sample for the SES was composed of 328 schools. It was composed of one nationally representative sample, randomly selected, and three purposive samples, selected to provide specific kinds of schools for comparative analyses. The four subsamples (described in detail in Hoepfner, Zagorski, and Wellisch, 1977) were as follows:

- 242 randomly selected representative schools,
- 14 schools that fed into or from the representative schools,
- 43 schools nominated as having promising compensatory programs, and
- 29 schools with high poverty but no compensatory funding.

The size of the representative sample was necessitated by the need to make highly accurate population projections about the state of the nation's compensatory education (CE) efforts. An even more critical need for the representative sample arose from the congressional mandate to ascertain the numbers of children who are and are not being served by Title I who are poor and low achieving. Arriving at accurate national projections to provide answers to Congress' questions required the large first-year sample.

After the first year of the study, descriptive projections to the nation were no longer planned. Instead, the study was to focus on inferential comparisons among various types, approaches to, and intensities of compensatory education in terms of their effectiveness. With the inferential goals in mind, it was necessary to maintain the three purposive samples in order to provide the variations needed for the analyses. But it was no longer necessary to maintain the large number of representative schools. The schools to be retained for the second year of the study were, of course, not to be atypical or nonrepresentative, the inferential requirements of the study simply assumed greater importance than the descriptive ones.

On this basis, 177 schools were retained for the second year (two schools left the study in the third year):

- 95 representative schools, selected primarily so that each cell of the first-year sampling matrix was represented by at least one school, but selected secondarily within cells with a systematic preference for schools with summer programs, with high percentages of low-achieving or poverty students, and with all six elementary-level grades;
- 13 schools that feed into or out of the representative schools selected for the second year;
- 41 schools nominated as having promising compensatory programs;
- 28 schools with high poverty but no compensatory funds.

THE REPRESENTATIVENESS OF THE REDUCED SAMPLE

After the selection of the reduced second-year sample, concerns were expressed that it might have become so unrepresentative, due to the selection criteria, that it would not be capable of supplying statistics that represent the nation's CE efforts. The first step in ascertaining whether this concern is justified is to inspect the distribution of the second-year sample over the sampling matrix for the first year. Because the sampling matrix provides cells that can be weighted to national projections, and has been found to provide highly accurate projections (Hoepfner et al., 1977), it is important to verify that the second-year sample has reasonable numbers of schools representing each of the strata, and does not have additional missing cells. (Exact proportionality of cell frequencies is not critical, however, because the comparative analyses merely need general representativeness, and weighting procedures could be used to adjust for disproportionalities if projections were to be made.)

The criteria for selecting the second-year sample guaranteed that no cells of the sampling matrix would be empty. Table 15-1 provides the unweighted distribution of schools according to the levels of each of the three sampling strata. It can be seen from Table 15-1 that each of the levels of each of the strata is well represented. The levels of the geographic region strata are still close to being proportionally represented, so that unweighted data would not have troublesome regional biases, and so that appropriately derived weights would provide a close approximation to the distribution of the population. The levels of the district size and district poverty strata are similarly represented. Table 15-1, however, presents information that merely assures us that the sample is not seriously distorted and would be capable of providing information that could be weighted, it does not indicate that the weighted information would necessarily be accurate estimates of the population.

The question remaining is whether the specifically selected sample can effectively support the comparative inferential investigations of the study. Clearly, inferential statistics such as t, F, or r are not as vulnerable to deviations from representative and random sampling as are population projections, but extremely biased sampling can have a biasing effect on them.

Table 15-1

Percentages of U.S. Public Schools With Grades in the 1-6 Range, by Strata and Level of the Sampling Matrix

Sampling Dimension	Population of the Nation's 62,534 Schools With Grades 1-6	Respondent Sample of 4,750 Schools	First-Year Representative Sample of 243 Schools	Second-Year Representative Sample of 95 Schools
Geographic Region				
New England	7.10	5.68	9.88	9.47
Metropolitan Northeast	8.32	7.4[illegible]	9.47	9.47
Mid-Atlantic	10.59	11.87	10.70	10.53
Southeast	14.58	20.72	10.29	10.53
North Midwest	17.53	16.23	11.93	10.53
South Central	15.47	13.56	11.11	9.47
Central Midwest	7.11	5 98	11.11	10.53
North Central	5.20	5.03	8.64	9.47
Pacific Southwest	9.87	10 00	9.88	12.63
Pacific Northwest	4.23	3.35	7.00	7.37
District Size				
Small District	45.51	33.24	36.21	35.79
Medium District	30.41	33.60	32.92	29.47
Large District	24.08	33.16	30.86	34.74
District Poverty Level				
High Poverty	22.05	33.41	29.22	28.42
Medium Poverty	29.96	32.63	34.57	34.74
Low Poverty	47.99	33.96	36.21	36.84

The effects of any nonrepresentativeness of the sample on the analytic outcomes, according to Kish (1965, pp 595-597) are not yet precisely identifiable, but there is a consensus that the inferential statistics planned for the comparative analyses are not seriously distorted by small deviations from true representativeness of the sample. Because we nonetheless did not want to push our luck in this matter, the second-year sample was selected in a manner that was expected to maintain representativeness while maximizing our control of important independent variables (e g , poverty level of school, availability of CE program by type and funding source).

THE ADEQUACY OF POPULATION PROJECTIONS BASED ON THE REDUCED SAMPLE

Logically, it is difficult to show that a sample is not unrepresentative in all important aspects. We elected to play the devil's advocate and test some projected statistics, even though we had no plans to use projected data for the comparative analyses. Our thinking went something like this: sampling weights will not make a bad sample look good—they are inferential aids, not cosmetics. Therefore, if we weight selected data (not the sampling dimensions of the previous section, which if weighted merely show whether we can calculate with accuracy) from the reduced sample and compare those projections to data in which we have confidence, we indirectly test how good the sample is. In other words, if the projections are close to what we expect, we are safe in inferring that the reduced sample is not seriously distorted.

Further, to determine precisely how accurate population projections based on the second-year sample actually are, it is necessary to make those projections based on newly calculated weights, but on data collected in the first year. In this way, the projections can be rigorously compared to known estimates with no contamination by history. Nine critical variables were selected as characteristics to be projected to test for accuracy. The nine characteristics are:

1. Urbanism of school
2. School and grade enrollments
3. Student poverty
4. Student minority concentration of school
5. Student race/ethnicity
6. School concentration of low-achieving students
7. Three-year stability (non-mobility)
8. Grade span type of school
9. School CE funding

The nine characteristics are projected from the 95 schools in the second-year representative sample and the projections are then compared to two other projected values. The critical comparison is with the projections from the Principal Sample of 4,750 schools surveyed during the planning year of the study. This sample was randomly selected and was sufficiently large so that error estimates were extremely small. The sample was used as a test of the accuracy of the projections made from the 243 representative schools (one of the schools later left the study, leaving 242 for the entire first year) in the first-year sample (Hoepfner et al., 1977). The projections from the first-year sample are also shown in the tables that follow for purposes of completeness. In addition, census information on race/ethnicity is also presented as another comparison to test for the accuracy of the second-year sample projections.

Urbanism of School. Schools were classified into four categories of population density and percentages of schools falling into each category were appropriately projected for each of the three samples. The result of the projections are presented in Table 15-2. The second-year sample's greatest projection error is 5.3 percent, an underestimate for small cities and rural areas near cities. An overestimate of 3.8 percent for medium-sized cities is in an adjacent category. Projections of percentages in large cities and in rural areas not near cities are accurate to less than .8 of one percent. If national projections were made from the reduced sample, we would expect them to be biased toward areas of higher population density. The distortion would, of course, be caused directly by the selection preferences for schools (see

Table VII-7 for the relationship between urbanism and poverty; Table VII-19 for the relationship with poverty; and Table VII-45 for the relationship with incidence of summer school; Hoepfner et al., 1977).

School Enrollment. Enrollment projections were made by enrollment categories and by absolute student counts. Table 15-3 presents the enrollment distribution by five categories of school size and Table 15-4 presents projected mean values of total school enrollment and of enrollment by grade.

In all five enrollment categories the second-year sample projections are within 1 percent of the projections of the Principal Sample. The second-year sample, in fact, appears to be better in this respect than the first-year sample from which it came.

Table 15-2

Projected Percentages of U.S. Public Schools With Grades in the 1-6 Range, by Urbanism of School

Urbanism	Principal Sample of 4,750 Schools	First-Year Sample of 243 Schools	Second-Year Sample 95 Schools
City of over 200,000	11.2	10.8	12.0
City of 50,000 to 200,000 or suburb	25.4	25.7	29.2
City under 50,000 or rural near city	38.5	36.4	33.2
Rural area not near city	24.9	27.1	25 6

Table 15-3

Projected Percentages of U.S. Public Schools With Grades in the 1-6 Range, by Enrollment of School

Enrollment in Grades 1-6	Principal Sample of 4,750 Schools	First-Year Sample of 243 Schools	Second-Year Sample of 95 Schools
1-200	28.4	34.1	29.4
201-400	40.4	40.1	40.6
401-600	21.4	18.5	21.5
601-800	7 0	5.2	6.7
801-2,800	2 7	2.0	1.8

From the projected enrollment numbers in Table 15-4, it can be seen that the second-year sample is composed of slightly larger schools than in either of the comparison samples. The larger enrollments are primarily at the primary grades, while the upper grades in the second-year sample are somewhat smaller (see Table VII-45 in Hoepfner et al., 1977, for the relation between enrollment and the availability of summer school).

Student Poverty. Three categories of poverty concentration of the schools were employed to test the projections of the second-year sample. From Table 15-5 it can be seen that the projections of the second-year sample over-estimate the high-poverty category by 7 percent and underestimate the lower poverty categories. These errors are directly expectable outcomes from the secondary selection criterion of poverty that was used for the second-year sample.

Table 15-4

Projected Percentages of U.S. Public Schools With Grades in the 1-6 Range

Enrollment Averages	Principal Sample of 4,750 Schools	First-Year Sample of 243 Schools	Second-Year Sample of 95 Schools
School Total	333 93	320 05	342 40
Grade 1	62 82	59 79	63 87
Grade 2	58 74	55 21	60 19
Grade 3	58 91	54 53	60 33
Grade 4	60 70	55 91	59 71
Grade 5	65 95	61 47	64 60
Grade 6	75 63	73 01	63 29

Table 15-5

Projected Percentages of U.S. Public Schools With Grades in the 1-6 Range, by Student Poverty Level

Percentage of Students Below Poverty	Principal Sample of 4,750 Schools	First-Year Sample of 243 Schools	Second-Year Sample of 95 Schools
0-20	51 6	50 0	46.5
21-50	32 3	31 7	30 4
51-100	16 1	18 3	23 1

Minority Concentration. Projections from the second-year sample of percentages of schools with each of four categories of minority concentrations are accurate to less than 2 percent The second-year sample projections tend to underestimate the low-concentration categories and to overestimate the high-concentration categories, clearly the result of the secondary selection criteria of poverty and low achievement used for the second-year sample.

Race/Ethnicity. On the basis of the projections from Table 15-6, it would be expected that some racial/ethnic groups will be overrepresented in projections from the second-year sample. In Table 15-7 it can be seen that the greatest overrepresentation is of the Black

Table 15-6

Projected Percentages of U.S. Public Schools With Grades in the 1-6 Range, by Minority Concentration

Percentage of Minority Students	Principal Sample of 4,750 Schools	First-Year Sample of 243 Schools	Second-Year Sample of 95 Schools
0-19	68.4	71.0	66.1
20-49	16.4	15.6	15.3
50-79	7.2	6.6	9.1
80-100	8.0	6.8	9.5

Table 15-7

Projected Percentages of Students Enrolled in U.S. Public Schools With Grades in the 1-6 Range, by Racial/Ethnic Group

Racial/Ethnic Group	HEW/OCR 1973 Directory Universe Projection*	Principal Sample of 4,750 Schools	First-Year Sample of 243 Schools	Second-Year Sample of 95 Schools
American Indian	0.52	0.91	0.93	0.77
Asian American	0.52	0.95	1.01	0.61
Black	15.22	16.17	13.55	18.44
Spanish Heritage	5.41	5.75	5.74	5.63
White	78.33	76.22	78.77	74.55

*This sample systematically excluded 8.3 percent of the public school students those who are associated with small school districts

282

percentage Percentages of other racial/ethnic groups are fairly close, being less than 2 percentage points away from the projections of the Principal Sample. The overrepresentation of minority students is clearly the result of the selection criteria of poverty and low achievement.

Concentration of Low-Achieving Students. The percentages of schools in three categories of percentages of students reading one grade or more below grade level were projected for the three samples, as reported in Table 15-8. The second-year sample projections overestimate the percentages of schools with larger percentages of low-achieving students. This finding is an indirect result of the secondary sample-selection criterion of low achievement.

Three-Year Stability Projections were made to three categories of school stability, defined by the percentage of students who remain in the school for three or more years, discounting matriculants and graduates. From Table 15-9, it can be seen that the second-year sample overestimates the percentage of the more stable (less mobility) schools by almost 5 percent.

Table 15-8

Projected Percentages of U.S. Public Schools With Grades in the 1-6 Range, by Concentration of Low-Achieving Students

Percentages of Students Reading Below Grade Level	Principal Sample of 4,750 Schools	First-Year Sample of 243 Schools	Second-Year Sample of 95 Schools
0-20	50.8	40.8	33.9
21-50	38.0	48.1	50.3
51-100	11.2	11.1	15.8

Table 15-9

Projected Percentages of U.S. Public Schools With Grades in the 1-6 Range, by Three-Year School Stability

Three-Year Stability (Non-Mobility) of Student Body	Principal Sample of 4,750 Schools	First-Year Sample of 243 Schools	Second-Year Sample of 95 Schools
Low (0-60% stay)	17.42	13.39	13.79
Medium (61-80% stay)	23.41	28.72	22.06
High (81-100% stay)	59.17	57.90	64.15

283

Grade-Span Type of School. Although schools embrace almost every conceivable combination of grades, four logical categories were defined and the percentages of schools in each category were projected from the three samples. The results in Table 15-10 indicate that the second-year sample overestimates the percentage of full six-grade schools, a direct result of that exact preference in the selection of the second-year sample.

CE Funding Four non-independent categories of schools were created according to whether or not they received certain kinds of CE funds. The comparison of percentages of schools in each category, presented in Table 15-11, indicates that the second-year sample overestimates the percentage of Title I participating schools by about 6 percent. This difference is probably not as large as it appears, due to the different years in which the data were collected and the recent trend for more schools to participate in specially funded compensatory programs.

Table 15-10

Projected Percentages of U.S. Public Schools With Grades in the 1-6 Range, by Grade-Span Type

Type of School by Grade Span	Principal Sample of 4,750 Schools	First-Year Sample of 243 Schools	Second-Year Sample of 95 Schools
Complete elementary	67.13	75 05	79.93
Feeder elementary	12 38	6 72	6 56
Primary	10 10	9 24	10.47
Intermediate/middle	10.39	8.99	2.58

Table 15-11

Projected Percentages of U.S. Public Schools With Grades in the 1-6 Range, by Compensatory-Education Funding Category

Category by Compensatory-Education Funding Source	Principal Sample of 4,750 Schools	First-Year Sample of 243 Schools	Second-Year Sample of 95 Schools
Title I funded	67 45	73 76	73.29
Other federal funds	28 33	29 73	27 64
State funded	41 82	44 70	44 29
District funded	15 52	20 38	19.70

CONCLUSIONS

Based on the comparisons of nine characteristics, the second-year sample provides no information that would lead us to expect serious or meaningful distortions in the comparative analyses. Furthermore, it is capable of providing data on which national projections could be based so that the projections rarely have errors above 5 percent (the non-randomness of the selection would, however, prevent our calculating unbiased standard errors). In most cases, the errors or projections are in the 1 and 2 percent range. It is concluded, therefore, that the second year sample can well be considered a representative sample and that study findings will not be significantly biased by its nature.

25:

CHAPTER 16. ATTRITION OF STUDENTS FROM THE SUSTAINING EFFECTS STUDY

Henry J. Zagorski
Lawrence A. Jordan
Edward J. Colon

Over a full calendar year, a stable core of 75 percent of the students in the Sustaining Effects Study (SES) remained enrolled in their schools, while the remaining 25 percent moved into or out of the schools. In order to determine if the attrition rate might be expected to bias the findings from the study, the rates were analyzed for important subgroups in the study. The stable students were found to have higher achievement scores than the attritors, so pure longitudinal samples will not be the same as a single-shot cross-sectional sample of the nation. The meanings of relational analyses of the study will probably not be affected by this, but attrition rates, considered in terms of additional student characteristics, are expected to have complex and recondite effects on comparisons among students grouped by their participation in CE programs.

Attrition is practically inevitable in a longitudinal investigation of the scope of SES. It is important to determine the extent to which attrition is related to student characteristics and analysis groups and whether attrition is likely to confound the analyses and interpretations of treatment effects of compensatory education. Campbell and Stanley (1966) indicate that attrition (experimental mortality) should be recognized as an important threat to study interpretability (internal validity). They state that "mortality, lost cases, and cases on which only partial data are available, are troublesome to handle and are commonly swept under the rug." This is especially true when such cases offer alternative explanations of observed educational-treatment gains. Anderson (1973) criticizes current longitudinal educational evaluation studies for seldom providing data on the number and type of dropouts occurring, and argues that attrition represents a major source of potential error in conclusions about the effects of educational treatments.

This study of attrition was conducted to assist readers in evaluating results and conclusions of the SES. If the students analyzed in the longitudinal investigation differ substantially from students in the total sample, both the internal validity of the study and the generalizability of the conclusions may be open to question. The primary objective of the investigation reported here was to examine how students in the longitudinal portion of the sample differ from students who either left prior to a full year's participation or entered the study late. The differences were also examined for subgroups defined by geographic region, district size, race/ethnicity, receipt of free or reduced-price meals, basic-skills achievement, and selection for compensatory education.

The data for the analysis came from seven major instrument files, providing a merged data file consisting of all students having data from the Comprehensive Tests of Basic Skills

(CTBS), Student Background Checklist (SBC), and Compensatory Education Roster (CER), from the Fall of 1976 to the Winter of 1977. During this period, the CTBS had been administered three times and the SBC and CER were each administered twice. From the merged file, students were selected who were present for at least one complete CTBS administration, were in grades 1 through 5 during the first year in the 95 schools constituting the Year 2 Representative Sample, and were not involuntary transfers between these schools and non-study schools (due to school grade-structuring within some districts). The total sample consisted of 35,808 students. The students were classified into six attrition groups described in Table 16-1, depending on the pattern of CTBS data they provided. The groups are coded in the left column of Table 16-1 with 'P' for 'Present' and 'A' for 'Absent' for each test administration in sequence.

TABULATIONS OF ACHIEVEMENT

The PPP, PPA, PAA attrition groups in Table 16-1 have scores on the reading and math achievement tests for the Fall, 1976 CTBS administration, while the PPP, APP, and AAP attrition groups have scores for the Fall, 1977 administration. The sixth group shown in Table 16-1 (APA) does not have data for either of these administrations, so it was omitted from the analysis of achievement scores. A series of three-group comparisons was made for each test occasion, basic skill, and cohort, using one-way ANOVA. Preliminary analysis on a 5 percent subsample showed that CTBS percentile scores, probit transformations of the percentiles, and vertical-scale scores all yielded essentially the same pattern of results. Therefore, the analyses presented here are in the form of the more easily interpreted percentile scores. Each ANOVA was supplemented by a Duncan Multiple Range Test to examine the significance of the three pair-wise group differences (1 vs. 2, 1 vs. 3, and 2 vs. 3). Results of the analysis are reported in Table 16-2. There are 20 ANOVAs represented in the table (2 test occasions x 2 basic skills x 5 cohorts). All 20 overall F ratios are significant at the .0001 level. Pair-wise group differences that are not significant at the .01 level by the Duncan Test are marked by connected 'bullets' (e.g., in cohort 4-5 for the first test administration, the reading and math pair-wise comparisons for the PPP and PPA groups are not significant).

An examination of the paired comparisons shows that students remaining in the study generally have higher achievement levels than the other students across subject areas, cohorts, and administrations. On the average, the PPP group exceeds the PPA group by 4.1 percentile points (this difference is .14 σ, where $\sigma = 28.8$, as based on the average within-group variance), the APP group by 5.3 percentile points (.18 σ), the AAP group by 7.5 percentile points (.26 σ), and the PAA group by 9.8 percentile points (.34 σ). Using Cohen's (1969) criterion for evaluating the size of mean differences, these obtained differences would be considered in the neighborhood of 'small' effects (.20 σ), and not large enough for 'medium' effects (.50 σ). Another commonly used criterion, although intended to be applied to program effects, is that differences of .33 σ can be considered 'educationally significant' (Horst, Tallmadge, and Wood, 1975) and, by this criterion, the PPP-PAA mean difference can be considered an educationally significant one. If treatment effects of about the same size are competing with these attrition effects, of course, the attrition effects may be very important as confounding factors for the analyses.

Table 16-1

Attrition Groups Used for Analysis

Group Code	Numbers of Students	Percentages of Students	Description
PPP	21,392	59.7	Present for each of the three test administrations. A small group of students that was absent from the second administration but present for both the first and third is also in this group, because their absence cannot be considered attrition from the study.
PPA	4,150	11 6	Absent for the third test administration only. This group is likely composed of students who transferred out between school years. However, some students who were still enrolled, but were absent for the third test administration in spite of make-up testing, are in this group.
PAA	3,115	8.7	Present for the first test administration but absent from the two subsequent ones. This group consists primarily of students who transferred out during the school year. However, it also includes students who were absent during the spring tests, then transferred to another school between school years.
APP	1,707	4.8	Absent for the first test administation but present for the following two. These could be students who were absent during the first administration in spite of make-up testing or students who transferred in during the regular school year.
AAP	4,238	11.8	Absent from both the first and second test administration, but present for the third. These are probably students who transferred in between school years.
APA	1,206	3.4	Present for only the second test administration. Consists primarily of students who transferred in during the regular school year, then transferred out again between school years. May also include new or old students with one or more absences in spite of make-up testing.
TOTALS	35,808	100.0%	

As shown in Table 16-2, the PAA group is significantly lower than the PPA group for all five cohorts. Although both groups include some students who did not transfer but were merely absent from the test administration, the PAA-PPA comparison strongly suggests that students who transfer out in the middle of a school year have lower achievement levels than those who transfer out between school years. In the right half of Table 16-2, the comparison of the two groups absent for the first and second test administrations indicates few consistent differences between students who transferred in during the school year (APP) and those who tansferred in between school years (AAP). Inspection of the connected 'bullets' in Table 16-2 also indicates that group differences are similar for reading and math scores, and are similar for all cohorts.

The calculations performed for Table 16-2 were repeated by proportionally weighting the students in each grade according to the SES sampling cell in which they reside. The resulting weighted means are presented in Table 16-3. The ANOVA F-ratios and Duncan

Table 16-2

Unweighted Comparisons of Reading and Math Percentile Means for Attrition Groups, by Cohort

Grade Cohort	Fall, 1976 Data: Group Code	N	Reading	Math	Fall, 1977 Data: Group Code	N	Reading	Math
1-2	PPP	4,885	48.1	48.6	PPP	4,885	52.2	53.2
	PPA	1,066	44.0	45.3	APP	474	48.0	48.4
	PAA	876	37.9	40.3	AAP	1,115	44.1	43.2
2-3	PPP	4,380	47.4	48.8	PPP	4,380	51.8	51.8
	PPA	979	43.5	44.0	APP	352	46.4	45.6
	PAA	702	38.1	38.6	AAP	991	41.6	41.9
3-4	PPP	4,506	46.6	47.4	PPP	4,506	51.1	50.7
	PPA	924	40.4	43.3	APP	358	47.0	47.2
	PAA	615	35.9	38.4	AAP	808	45.6	44.8
4-5	PPP	4,138	47.2	47.9	PPP	4,138	50.1	48.3
	PPA	706	45.8	46.4	APP	303	44.0	43.6
	PAA	534	37.2	40.7	AAP	789	44.8	44.8
5-6	PPP	3,483	49.6	50.3	PPP	3,483	52.9	53.0
	PPA	475	44.1	44.1	APP	220	43.7	45.3
	PAA	388	37.9	37.9	AAP	535	45.3	45.6

Note Connected bullets indicate non-significant pair-wise differences of the means Numbers of students are the same across both administrations only for the PPP group Other groups and totals, of course, are composed of different numbers of cases

Tests are not applicable here since, in effect, the table reflects 'population estimates' and the mean square error calculations are not meaningful. Strictly speaking, the use of the total grade sample weights to compute mean estimates for subpopulations within grade is not legitimate. However, the procedure was not employed with the aim of making grade subpopulation estimates. Rather, it was simply adopted as a device to examine how the weights and the percentile scores would interact to modify the observed unweighted differences between the attrition groups.

Comparison of the weighted group means in Table 16-3 with the unweighted ones in Table 16-2 across all test occasions, subjects, and cohorts shows that the PPP group increased by an average of 2.9 percentile points, the PPA group by 1.2 percentile points, and the PAA group by 2.3 percentile points, while the APP and AAP groups decreased slightly. Thus, there is a slight tendency for the overall mean differences between the PPP group and the other groups to increase when weighting is applied.

Table 16-3

Weighted Comparisons of Reading and Math Percentile Means for Attrition Groups, by Cohort

Grade Cohort	Fall, 1976 Data			Fall, 1977 Data		
	Group Code	Reading	Math	Group Code	Reading	Math
1-2	PPP	50 7	51 3	PPP	55 3	55.0
	PPA	44 6	47 1	APP	47 7	48 4
	PAA	39 9	41 8	AAP	45 5	45.7
2-3	PPP	49.6	52 2	PPP	55 7	56.6
	PPA	44 2	47 4	APP	47 5	48 2
	PAA	37 1	39.2	AAP	43 1	44.2
3-4	PPP	49 6	50.2	PPP	54.3	54 3
	PPA	43.0	45 7	APP	45.6	47 4
	PAA	38.7	40.3	AAP	43.4	41 9
4-5	PPP	50 4	50.4	PPP	53.4	51.8
	PPA	46 1	47 2	APP	45.3	43.9
	PAA	40 2	43 4	AAP	43.4	43.4
5-6	PPP	51 4	52 9	PPP	54 4	55 8
	PPA	42 2	45.0	APP	43.9	37 1
	PAA	43 1	41 9	AAP	43 8	44.6

TABULATIONS OF STUDENT BACKGROUND AND OTHER CHARACTERISTICS

The six attrition groups shown in Table 16-1 were combined into three attrition categories, as follows:

Attrition Category	Attrition Groups
Departed	PPA, PAA, APA
Entered	APP, AAP
Retained	PPP

The 35,808 students (all cohorts combined) were cross-tabulated using the attrition category as one dimension, and region, district size, basic-skills achievement level, race/ethnicity, receipt of free or reduced-price meals, and reading and math CE participation, in turn, as the second dimension. The results appear in Table 16-4.

There were small amounts of missing data for some of the characteristics reported in Table 16-4, due to incomplete responses. The cases could be assigned to attrition groups based on the CTBS, but they were missing information on the SBC and CER. Accordingly, for each missing SBC and CER characteristic, the cells within each column were adjusted by distributing the cases with missing data in the same proportions as the cases that had data available. Row percentages were then computed using the adjusted cell frequencies. The chi-square values computed for both the adjusted and unadjusted tables were all significant at the .0001 level. All the tabulations involving reading and math CE participation were further adjusted to counteract the effects of school reporting errors on the CER for CE students who left the study during the school year. These errors (ambiguities in how school coordinators categorized students who had already departed) resulted in numerous CE students being reported as non-CE students, which artificially inflated the retention rates for the program categories and deflated them for the non-program categories. (The adjustment consisted of converting the CE status data for all students who left the study during the school year (PAA group) into missing value indicators, and then prorating the tabulations in the same manner as that described above for students having no CER data.)

The retention rates in Table 16-4 are combined and summarized in Table 16-5, which shows the percentages of students in the 'Retained' category within several classifications of student characteristics, ranked in order from highest to lowest. The percentages in Table 16-5 are not independent of one another, because students in one set of characteristics are also included in all other sets. However, the general trend of the findings is still clearly visible. Students in small districts have the highest retention rates (71.2 percent), while students residing in the Pacific region have the lowest retention rates (50.1 percent). As expected, students in large districts, students with low achievement, recipients of free or reduced-price meals, and minority students all tend to have low retention rates. Other-CE participants in Title I schools also show a relatively low retention rate. Students in the Midwest, Southeast, and Northeast, regular achievers, and non-recipients of free or reduced-price meals show relatively high retention rates.

Table 16-4

Percentages of Students in the Representative Sample by Attrition Category and by Region, District Size, Basic-Skills Achievement Level, Race/Ethnicity, Receipt of Free or Reduced-Price Meals, and Reading and Math CE Status

Characteristic	Attrition Category: Departed	Entered	Retained	Total Count
Region				
Northeast	21.5	14.6	63.9	11,207
Southeast	21.7	13.4	64.9	4,255
Midwest	22.1	13.0	64.9	3,480
Central	21.9	20.2	57.9	9,818
Pacific	31.4	18.5	50.1	7,048
District Size				
Small	16.1	12.7	71.2	9,078
Medium	25.0	17.5	57.5	9,983
Large	27.0	18.1	54.9	16,747
Basic-Skills Achievement Level				
Low (at or below 35th percentile)	28.1	17.9	54.0	14,560
Regular (above 35th percentile)	20.6	15.7	63.7	21,248
Race/Ethnicity				
Black	25.6	16.9	57.5	7,867
Hispanic	28.7	18.1	53.2	3,861
Other	22.2	16.3	61.5	24,080
Free or Reduced-Price Meals (poverty)				
Receive	27.9	17.7	54.4	15,320
Do not receive	20.5	15.8	63.7	20,488
Reading CE Status; Students Receive				
Title I plus other CE	29.4	8.1	62.5	706
Title I only	23.9	17.0	59.1	5,195
Other CE in Title I schools	31.5	13.6	54.9	2,334
Other CE in other-CE schools	28.0	16.1	55.9	1,348
No CE in Title I schools	20.3	19.0	60.7	16,528
No CE in other-CE schools	26.0	12.2	61.8	7,048
No CE in non-CE schools	27.3	17.4	55.3	2,649
Math CE Status; Students Receive				
Title I plus other CE	27.1	14.6	58.3	362
Title I only	25.0	17.1	57.9	3,369
Other CE in Title I schools	32.2	13.7	54.1	1,753
Other CE in other-CE schools	25.5	18.4	56.1	885
No CE in Title I schools	20.9	18.4	60.7	19,280
No CE in other-CE schools	26.5	12.1	61.4	7,510
No CE in non-CE schools	27.3	17.4	55.3	2,649
Total Count	8,471	5,945	21,932	35,808

LOG-LINEAR ANALYSES OF THE JOINT EFFECTS OF ACHIEVEMENT, POVERTY, ETHNICITY, URBANISM, AND REGION ON RETENTION RATES

The MULTIQUAL program (Bock and Yates, 1973) was used to model retention rates as a function of the classifying variables, in order to assess the joint or combined effects of the variables. As shown in Table 16-5, students in small districts and students in the Midwest, Southeast and Northeast regions tend to have high retention rates. It is of interest to know whether Urbanism and Region are independently related to retention or whether these particular regions have higher retention rates because they also tend to have smaller school districts. Similarly, Table 16-5 shows that Hispanic students and students in the Pacific region tend to have low retention rates. If ethnicity and region are independently related to

Table 16-5

Rank Ordering of Classifications of Students, by the Percentages of Students Retained

Rank	Student Classifications	Percentages Retained
1	Students in small districts	71.2
2	Students in the Midwest	64.9
3	Students in the Southeast	64 9
4	Students in the Northeast	63 9
5	Students not receiving free or reduced-price meals	63 7
6	Students achieving at regular levels	63.7
7	Students not selected for CE services in other-CE schools	61.6
8	White and non-black or non-Hispanic students	61 5
9	Students selected for Title I and other-CE services	61 1
10	Students not selected for CE services in Title I schools	60 7
11	Students receiving CE services from Title I only	58 6
12	Students in the Central U S	57 9
13	Students in medium-sized districts	57.5
14	Black students	57.5
15	Students selected for CE services in other-CE schools	56 0
16	Students in schools with no CE	55 3
17	Students in large districts	54.9
18	Students selected for other-CE services in Title I schools	54.6
19	Students receiving free or reduced-price meals	54 4
20	Students achieving at low levels	54.0
21	Hispanic students	53.2
22	Students in the Pacific region of the U S	50.1

retention, then we would predict that Hispanic students in the Pacific region would have particularly low retention rates. On the other hand, since almost exactly half of the Hispanic students are from the Pacific region (49.9 percent), we could find by an analysis of joint effects that the Pacific region has low retention rates because of a disproportionate number of Hispanic students or, if region is the primary factor, that Hispanic students have low retention rates because they happen to come disproportionately from the Pacific region. When we look at ethnicity and region together, as shown below, we find that none of these results is obtained. In fact, Hispanic students in the Pacific region tend to have rather average retention rates, and we must look elsewhere for a basis of the low retention rates among Hispanic and Pacific region students. Thus, it is important to assess the joint effects of variables, since an examination of variables one at a time can be misleading.

Our approach to assess the joint effects of the variables is by means of an *N*-way table of frequencies for students retained and not retained, by *N*-1 classifying variables. Retention rate is then considered the 'response factor' or dependent variable in a log-linear model, and the *N*-1 other variables are considered the 'subject factors' or independent variables.

This approach does have some disadvantages. It is not practical to model all of the classifying variables simultaneously, because there are limits to the size of problems handled by the MULTIQUAL program, and because cell sizes become quite small and the cell retention rates become correspondingly unstable with too many classifying variables. Accordingly, we attempted to examine a number of cross-classifications having three or four classifying variables For a first analysis of joint effects, we examined five tables having the form: VARIABLE x COHORT x CE-GROUP x RETENTION, substituting the variables ACHIEVEMENT, LOW INCOME, ETHNICITY, URBANISM and REGION, in turn, for the VARIABLE term in the above model.

Throughout the remainder of this section, we will capitalize the names of qualitative variables, which will have categories defined as follows:

ACHIEVEMENT — *Low*—Scored at 33rd Percentile or below on Total CTBS at the first recorded administration. *Regular*—Otherwise.

LOW INCOME: *Poor*—Recipient of free or reduced-price meals. *Non-Poor*—Not a recipient. *Missing*—SBC (basis for this classification) not available.

ETHNICITY: *Black*, *Hispanic*, or *White/Other*—as shown on SBC. *Missing*—SBC not available.

URBANISM — *Small*, *Medium*, or *Large* district size—a surrogate available from sampling frame.

REGION — *Northeast* (Standard Federal Regions I, II, III), *Southeast* (IV), *Midwest* (V), *Central* (VI, VII, VIII), or *Pacific* (IX, X)—available from sampling frame.

CE-GROUP: *Title I, Other CE,* or *Non-CE*—Participants in Title I, Other-CE, or Non-CE programs, respectively. (This variable was defined only in terms of Reading CE, since it was felt that Math CE would yield essentially identical results.)

RETENTION: *Yes*—Student has valid CTBS data for fall test administrations (PPP group in Table 1). *No*—Otherwise.

The variables and their categories should be mostly self-explanatory. Note that the students with missing data for LOW INCOME are also missing data for ETHNICITY, since these variables were obtained from the same instrument. The REGION variable was collapsed from ten categories to five categories, with the aim of obtaining fewer regions having more equal population sizes; ten categories were felt to be too many to handle analytically using MULTIQUAL. The ETHNICITY and CE-GROUP categories have been similarly collapsed. Finally, the two RETENTION categories contrast the PPP STUDENTS (retained) with all others (not retained). The log-linear model analyzes retention rates as a function of the classifying variables or, more precisely, they model *logits* having a one-to-one relationship with the Retention rates (½ ℓn [Retained/Not Retained]) as a function of the classifying variables.

ACHIEVEMENT, COHORT, AND CE-GROUP

The ACHIEVEMENT x COHORT x CE-GROUP x RETENTION analysis is typical, and will be discussed in detail in order to illustrate the method.

Table 16-6 summarizes the log-linear analysis of these variables. In the language of the MULTIQUAL program, ACHIEVEMENT, COHORT and CE-GROUP were considered as 'subject factors,' and RETENTION as the 'response factor.' All 35,808 students were classified into one of the cells of the resulting table, and for this table we obtained an overall $\chi^2(30) = 604.66$ ($p < .001$), indicating wide variations in the retention rates for the cells. (If all 30 cells—two levels of ACHIEVEMENT, by five levels of COHORT, by three levels of CE-GROUP—had identical retention rates, this χ^2 would have been zero. The high χ^2 value indicates a poor fit of the model requiring all cells to have identical rates.)

As shown in the top panel of Table 16-6, all but 8.9 percent of the variation in retention rates is accounted for by the three main effects, and all but 0.9 percent by the main effects and the three, two-way interactions. Thus, we can predict retention rates quite well as a function of the main effects, but further precision is provided by including one or more of the two-way interactions. The three-way interaction is not needed, since $\chi^2(8) = 5.69$, which is not significant, after the two-way interactions are entered.

When we attempt to isolate the main effects and interactions that are most strongly related to retention rates, we have a minor problem, which is that the results depend to some extent on the order in which the classifying variables are entered. As in a stepwise multiple regression analysis, the classifying variables tend to account for more of the variation in

Table 16-6

Log-Linear Analysis of Retention as a Function of ACHIEVEMENT, COHORT, and CE-GROUP

Effect	X^2	df	Probability	Percentage of Total X^2
	Total X^2			
Constant	604.66	29	< 001	100.0
All Main Effects	53.78	22	< .001	8.9
All 2-Way Interactions	5.69	8	682	0.9
Full Model	0.00	0	—	0 0
	Partition of X^2			
ACHIEVEMENT	314.47-334.38	1	001	52 0-55.3
COHORT	189 36-202 50	4	.001	31 3-33.5
CE-GROUP	22 96- 42.35	2	001	3.8- 7.0
ACHIEVEMENT x COHORT	0.86- 3.69	4	.930-.450	0 1- 0 6
ACHIEVEMENT x CE-GROUP	29.55- 30 90	2	001	4.9- 5.1
COHORT x CE-GROUP	13.75- 17.45	8	.089-.026	2.3- 2.9
ACHIEVEMENT x COHORT x CE-GROUP	5 69	8		0.9

rates when they are entered into the model at an early step than when they are entered at a late step. In the lower panel of Table 16-6, ranges of observed X^2 values are reported for all possible orders of main effects and then controlling for main effects, for all possible orders of two-way interactions. These ranges are usually narrow, and presumably bracket the 'true' X^2 values. As main effects, ACHIEVEMENT, COHORT and CE-GROUP all account for significant variation in retention rates, with ACHIEVEMENT and COHORT together accounting for over 80 percent of the variation, and CE-GROUP accounting for a much smaller percentage. There is also a significant ACHIEVEMENT x CE-GROUP interaction, and borderline COHORT x CE-GROUP interaction. The ACHIEVEMENT x COHORT Interaction is the smallest effect, and is clearly not significant.

Table 16-7 displays the marginal tables of retention rates for all significant or near-significant terms in the model. ACHIEVEMENT accounts for the largest differences in retention rates, with a 10 percent difference between low and regular achievers. COHORT is also an important differentiator of retention rates, which increase in a linear manner from 56 percent for Cohort 1-2 to 66 percent for Cohort 5-6. Thus, low achievers and students in the lower cohorts tend to have lower retention rates than regular achievers and students in the higher cohorts. The CE-GROUP effect is smaller than the other two main effects, and shows that the other-CE group has lower retention than the other two groups, and that Title 1 group has slightly lower retention (2 percent lower) than the non-CE group.

Table 16-7

Marginal Values of Retention Rates for Significant Effects in the ACHIEVEMENT x COHORT x CE-GROUP x RETENTION Analysis of Table 16-6

		Low	Regular	
ACHIEVEMENT		54	64	

	1-2	2-3	3-4	4-5	5-6
COHORT	56	57	.60	.62	.66

	Title I	Other CE	Non-CE
CE-GROUP	59	.55	.61

		Title I	Other CE	Non-CE
ACHIEVEMENT x CE-GROUP	Low	57	54	52
	Regular	64	56	64
	(Difference)	(07)	(02)	(12)

		Title I	Other CE	Non-CE
COHORT x CE-GROUP	1-2	54	52	.57
	2-3	54	52	.58
	3-4	62	55	61
	4-5	63	61	.62
	5-6	64	61	.67

The result for the main effects must be qualified, however, by the results for the ACHIEVEMENT x CE-GROUP interaction. ACHIEVEMENT is associated with only a 2 percent difference in retention for the other-CE students, but a 7 percent difference for Title I students, and a 12 percent difference for non-CE students. This interaction has important implications for an analysis of achievement gains in the longitudinal study. In a comparison of Title I and non-CE students, for example, we will find that the Title I group has a higher retention rate for low achievers that the non-CE group. This differential retention rate can be considered to be a benefit of Title I programs, because they are apparently more successful in keeping low achievers in school, but it may also have the inadvertent effect of making Title I seem academically harmful since there are more low achievers available for retesting in the Title I group This would be true, of course, if the opposite effects of regression did not operate and if the comparison group truly represented its national population (but see the article on the reduced sample in this volume that indicates that our comparison groups are slightly more biased towards low achievers than the Title I group). Thus, if Title I and non-CE programs had the same effect on achievement, the non-CE group would tend to show higher achievement gains, solely as a function of differential retention rates and lower gains as a function of regression to the mean, but would not be much influenced by the sampling. We will return to this point again in the conclusion.

The COHORT x CE-GROUP effect in Table 16-7 can be described as follows: within each of the CE groups, the retention rates increase for the higher cohorts, with 'jumps' or larger shifts in the size of retention rates at the points indicated by horizontal lines; and the COHORT x CE-GROUP interaction reflects the fact that these 'jumps' take place at different points for different CE groups. This interaction should probably not be taken seriously, however, since the associated χ^2 is only marginally significant, and does not reach significance at the .05 level when the other two-way interaction terms are entered first. This means that the effect would probably not be obtained, or would not be obtained in the same form, on replication.

ACHIEVEMENT, ETHNICITY, LOW INCOME, CE-GROUP, AND RETENTION

When we repeated the analysis of the last section, with LOW INCOME replacing ACHIEVEMENT, the resulting tables were very similar, apart from the influence of the Missing-Data category for LOW INCOME. Students missing SBC data (about 5 percent of the total) were overwhelmingly from the PAA group, so these students were present for the fall CTBS administration and then left before other kinds of data were collected. We have no way of assigning valid LOW INCOME or ETHNICITY categories to these students, but can assume that they have relatively high proportions of free or reduced-price lunch recipients and minority students. We do have information about the relationship between ACHIEVEMENT and LOW INCOME, as shown in Table 16-8.

Thus, more low achievers than regular achievers are missing information on the LOW INCOME variable, even though only 41 percent of students are classified as low achievers. There is also a fairly strong association between LOW INCOME and ACHIEVEMENT (with

Table 16-8

Cross-Tabulations of Students by ACHIEVEMENT and LOW INCOME Categories

		LOW INCOME			
		Poor	Non-Poor	Missing	Total
ACHIEVEMENT	Low	8,791	4,770	1,008	14,569
	Regular	5,628	14,775	836	21,239
	Percent Low	61%	24%	55%	41%

the Missing column omitted, ϕ = .349 for the remaining 2 x 2 table), so it is not surprising that the two variables have similar relationships with RETENTION.

As shown in Table 16-9, poor students had lower retention rates (58 percent) than non-poor (67 percent), overall, but this result must be qualified by the results from the LOW INCOME x CE-GROUP table, which indicate a 3 percent difference between poor and non-poor in the other-CE group, a 6 percent difference in the Title I group and a 10 percent difference in the non-CE group. This is almost identical to the pattern obtained for the ACHIEVEMENT x CE-GROUP interaction in Table 16-7.

Table 16-9

Selected Marginal Tables of Retention Rates From a Log-Linear Analysis of LOW INCOME, COHORT, CE-GROUP and RETENTION

		Poor	Non-Poor	Missing
LOW INCOME		58	.67	.01
LOW INCOME x CE-GROUP	Title I	60	66	01
	Other CE	57	60	02
	Non-CE	57	67	.01

These results for LOW INCOME and similar results for the ETHNICITY variable led to an attempt at predicting retention rates from ACHIEVEMENT, ETHNICITY, LOW INCOME and CE-GROUP simultaneously. For this analysis, students in the Missing category for LOW INCOME and ETHNICITY were deleted, and the data were collapsed across COHORT. The statistical summary for this analysis is given in Table 16-10.

The results from this analysis are not simple to interpret. Table 16-10 indicates that a significant variation in rates is still present after all of the three-way interactions have been entered. Over 96 percent of the variation in rates has been accounted for at that point, and further fitting would probably not yield interpretable results. In the lower panel of Table 16-10, we can identify effects worthy of further examination, by selecting the effects that are significant at the .05 level regardless of order of entry. By this criterion, we select all four main effects, three of the two-way interactions, and two of the three-way interactions. The model containing only these effects yields $\chi^2(15) = 32.83$, $p = .005$, and may be regarded as fitting reasonably well. The marginal values of retention rates for these effects are given in Table 16-11.

In Table 16-11, we can make certain generalizations about retention rates from the main-effect marginals:

- Regular achievers have higher rates than low achievers.

Table 16-10

Log-Linear Analysis of RETENTION as Function of ACHIEVEMENT, ETHNICITY, LOW INCOME and CE-GROUP

Effect	χ^2	df	Probability	Percentage of Total χ^2
	Total χ^2			
Constant	585 81	35	< 001	100 0
All Main Effects	159 62	29	<.001	27 2
All 2-Way Interactions	52 87	16	< 001	9 0
All 3-Way Interactions	19 78	4	< 001	3 3
	Partition of χ^2			
ACHIEVEMENT	106 72-241 57	1	< 001	18 2-41 2
ETHNICITY	13 50-104 89	2	002-< 001	2 3-17 9
LOW INCOME	125 54-282 67	1	< 001	21 4-48 3
CE GROUP	18 06- 55 72	2	< 001	3 1- 9 5
ACHIEVEMENT x ETHNICITY	1 35- 4 39	2	512- 111	0 2- 0 8
ACHIEVEMENT x LOW INCOME	0 03- 1 29	1	862- 256	0 0- 0 2
ACHIEVEMENT x CE-GROUP	9 38- 14 72	2	009-< 001	1 6- 2 5
ETHNICITY x LOW INCOME	1 67- 6 47	2	434- 039	0 3- 1 1
ETHNICITY x CE-GROUP	20 64- 71 28	4	< 001	3 5-12 2
LOW INCOME x CE-GROUP	16 45-70 70	2	< 001	2 8-12 1
ACHIEVEMENT x ETHNICITY x LOW INCOME	12 66- 15 67	2	002-< 001	2 2- 2 7
ACHIEVEMENT x ETHNICITY x CE-GROUP	10 95- 12 33	4	027- 015	1 9- 2.1
ACHIEVEMENT x LOW INCOME x CE-GROUP	1 62- 3 13	2	.445- 209	0 3- 0.5
ETHNICITY x LOW INCOME x CE-GROUP	4 72- 9 67	4	317- 046	0.8- 1.1

- White/other students have higher rates than black students, and black students have higher rates than Hispanic students.
- Non-poor students have higher rates than poor ones.
- Non-CE students have slightly higher rates than Title I students, and Title I students have higher rates than other-CE students.

The ACHIEVEMENT and LOW INCOME main effects are the strongest, and together they account for over half of the variation in rates.

As the interactions show, however, almost all of the above generalizations fail to hold for one or more subgroups.

Most of the interactions in Table 16-11 involve the CE-GROUP variable, and especially the

Table 16-11

Marginal Values of Retention Rates for Significant Effects in the ACHIEVEMENT x ETHNICITY x LOW INCOME x CE-GROUP x RETENTION Analysis of Table 16-10

ACHIEVEMENT	Low	Regular
	58	66

ETHNICITY	Black	Hispanic	White/Other
	61	57	65

LOW INCOME	Poor	Non-Poor
	58	67

CE GROUP	Title I	Other CE	Non-CE
	62	58	64

ACHIEVEMENT x CE-GROUP		Title I	Other CE	Non-CE
	Low	60	57	57
	Regular	66	59	67

ETHNICITY x CE-GROUP		Title I	Other CE	Non-CE
	Black	62	47	62
	Hispanic	57	61	54
	White/Other	63	60	.65

LOW INCOME x CE-GROUP		Title I	Other CE	Non-CE
	Poor	60	57	57
	Non-Poor	65	59	67

ACHIEVEMENT x ETHNICITY x LOW INCOME	Black			Hispanic			White/Other		
	Low	Reg	(Diff)	Low	Reg	(Diff)	Low	Reg	(Diff)
Poor	58	61	(02)	55	57	(02)	53	68	(.09)
Non-Poor	60	73	(13)	55	64	(09)	63	68	(05)
(Poor Non-Poor)	(–/02)	(– 12)	(– 11)	(00)	(– 07)	(– 07)	(– 10)	(– 06)	(04)

ACHIEVEMENT x ETHNICITY x CE-GROUP	Black			Hispanic			White/Other		
	Low	Reg	(Diff)	Low	Reg	(Diff)	Low	Reg	(Diff)
Title I	63	61	(– 02)	55	66	(11)	60	68	(.08)
Other CE	47	46	(– 01)	61	62	(01)	58	62	(04)
Non-CE	58	68	(10)	51	58	(07)	58	67	(09)
(Title I-Non-CE)	(05)	(– 07)	(– 12)	(04)	(08)	(04)	(-02)	(-01)	(– 01)

301

Other-CE subgroup. As we have seen, the other-CE students tend to have low retention rates, and do not show the usual difference in rate between low and regular achievers, or between poor and non-poor students. When we examine the ETHNICITY x CE-GROUP interaction, however, we discover that the low rate for other-CE students is primarily due to the very low rate for the black subgroup (47 percent), and that the Hispanic subgroup actually has its highest retention rate for other-CE students (61 percent. This probably indicates that other-CE programs serving black students are different from the other-CE programs serving Hispanic students. As shown in Table 16-12, it is also the case that white/other and black students are more likely to receive Title I funds, while the reverse is true for the Hispanic group (but a greater *proportion* of Hispanic students participate in Title I than either of the other groups).

Table 16-12

Cell Frequencies for ETHNICITY x CE-GROUP Cross-Classification

	Title I	Other CE	Non-CE
Black	1,963	708	4,777
Hispanic	884	978	1,775
White Other	2,851	1,844	18,214

These results suggest the need for caution in contrasting students receiving other-CE funds with other students, since the other-CE subgroup is somewhat heterogeneous. We cannot say what role bilingual programs (technically not usually defined as 'compensatory,' but frequently considered as such by school personnel) play in these findings.

Three-way interactions are always difficult to characterize, and Table 16-11 includes two of them. The ACHIEVEMENT x ETHNICITY x CE-GROUP subtable at the bottom of Table 16-11 is particularly interesting. The main effect for CE-GROUP indicates that the retention rate for non-CE students (64 percent) is slightly higher than the rate for Title I students (62 percent). Examination of the ACHIEVEMENT x ETHNICITY x CE-GROUP marginals will show, however, that non-CE students have lower retention rates for five of six subgroups, and only black-regular achieving students show a higher retention rate for non-CE (68 percent) than for Title I students (61 percent). How can that be? The apparent paradox is resolved if the retention rates and cell frequencies are examined simultaneously, as in Table 16-13.

Thus, the overall rate for non-CE students is pulled up by the large group of white/other regular achievers, even though non-CE students have *lower* retention rates in five of six columns of Table 16-13.

We noted earlier that Title I programs retain more low achievers than non-CE programs, resulting in a bias toward making Title I look bad. Further complications ensue if we use the

Table 16-13

Retention Rates and Cell Frequencies for ACHIEVEMENT x ETHNICITY x CE-GROUP Subtable

	ETHNICITY						
	Black		Hispanic		White/Other		Totals
CE-GROUP	Low	Reg.	Low	Reg.	Low	Reg.	
Title I	63 1,641	.61 322	.55 721	.66 164	.60 1,860	68 990	.68 5,698
Other CE	.47 469	.46 239	.61 686	.62 294	58 951	.62 892	.58 3,531
Non-CE	.58 2,626	68 2,151	.51 1,004	58 772	.58 3,614	.67 14,598	.64 24,765
Totals	.58 4,736	65 2,712	55 2,411	.60 1,230	.59 6,425	.67 16,480	

data of Table 16-13 to compare Title I and non-CE programs within ethnic subgroups. There are three different results:

- *White/Other Students.* Title I retains slightly more low achievers than non-CE (2 percent difference), and also slightly more regular achievers (1 percent difference), resulting in a slight but probably trivial bias toward making Title I look bad.
- *Hispanic Students.* Title I retains more low achievers than non-CE (4 percent difference), but retains even more regular achievers (8 percent difference), resulting in a net bias toward making Title I look good.
- *Black Students.* Title I retains noticeably more low achievers than non-CE (5 percent), and retains many fewer regular achievers (7 percent difference in the other direction), resulting in a net bias toward making Title I look bad.

Thus, an analysis of achievement gains by ethnic subgroups could have quite mischievous effects, showing no difference or gains for Title I, or losses for Title I, solely as a function of differential attrition, and depending on the subgroup analyzed.

REGION, URBANISM, CE-GROUP, AND RETENTION

REGION and URBANISM (i.e., size of school district) are both variables from the sampling frame for the Sustaining Effects Study. We attemped to determine their joint effects, along with CE-GROUP, on RETENTION. The statistical summary is given in Table 16-14, and the relevant marginal values are given in Table 16-15.

Table 16-14

Log-Linear Analysis of RETENTION as a Function of REGION, URBANISM, and CE-GROUP

Effect	χ^2	df	Probability	Percentage of Total χ^2
	Total χ^2			
Constant	1,443.63	44	<.001	100.0
All Main Effects	340.05	38	<.001	23.6
All 2-Way Interactions	81.26	16	<.001	5.6
	Partition of χ^2			
REGION	408.26-444.86	4	<.001	28 3-30.8
URBANISM	653.72-685.25	2	<.001	45.3-47 5
CE-GROUP	0.19- 34.88	2	909-<.001	00.0--2.4
REGION x URBANISM	130 75-136.92	8	<.001	9.1- 9.5
REGION x CE-GROUP	80.26- 98.11	8	<.001	5.6- 6.8
URBANISM x CE-GROUP	24 85- 41.61	4	<.001	1.7- 2.9
REGION x URBANISM x CE-GROUP	81.26	16	<.001	5.6

The largest effect is for URBANISM, which accounts for nearly half of the variation in rates. Small districts have retention rates that are roughly 15 percent higher than medium and large districts, and a pure longitudinal sample will have a disproportionate number of students from small districts. REGION accounts for more than a quarter of the variation in rates, and reveals an east/west contrast, with relatively high rates (64-65 percent) in the Northeast, Southeast and Midwest regions, and lower rates (58 percent and 50 percent) in the Central and Pacific regions. As shown by the REGION x URBANISM marginals in Table 16-15, students from small districts in the East tend to have the highest rates, and students from large districts in the Central and Pacific regions tend to have the lowest rates, which is what we would expect if REGION and URBANISM were independently related to retention. There is a substantial REGION x URBANISM interaction, however, which accounts for about 9 percent of the total variation, and which appears to be attributable to atypical values for the Midwest-Medium cell. By taking simple deviations from the marginal rates, we obtain the following table of REGION x URBANISM deviations:

	Small	Medium	Large
Northeast	.03	−.01	−.01
Southeast	−.01	.00	−.04
Midwest	−.04	.24	−.02
Central	−.04	−.01	.02
Pacific	.01	−.01	.02

Table 16-15

Marginal Values of Retention Rates for Significant Effects in the REGION x URBANISM x CE-GROUP x RETENTION Analysis of Table 16-14

		Small	Medium	Large
URBANISM		.71	.57	.55

		Title I	Other CE	Non-CE
CE-GROUP		.60	.55	.60

		Northeast	Southeast	Midwest	Central	Pacific
REGION		64	65	.65	58	.50
REGION x CE-GROUP	Title I	61	.61	71	62	.51
	Other CE	.59	.61	57	.50	.53
	Non-CE	65	66	.65	.58	.48
REGION x URBANISM	Small	78	.75	.72	.65	.63
	Medium	61	.62	.76	54	47
	Large	58	56	.58	.55	.47

		Small	Medium	Large
URBANISM x CE-GROUP	Title I	70	.59	.53
	Other CE	.64	.51	.55
	Non-CE	.72	58	.55

Thus, the rate deviation for the Midwest-Medium cell stands out since it is 24 points higher than expected. In Table 16-15, the rate for the Midwest-Medium cell more closely resembles the rate for the Midwest-Small cell, while the rate for medium districts in other regions resembles the corresponding rate for large districts. In the Midwest region, the schools in the medium subgroup are from larger towns in predominantly rural areas, and have a more rural character than their district sizes would imply. Students and their families may be more mobile in areas served by medium and large school districts, partly explaining their lower retention rates. The consequences of parental mobility for attrition are also greater in suburban and urban areas, since a move of only a few miles within a metropolitan area is likely to mean a transfer to another school.

All of the two-way interactions in Table 16-14 are significant, but the CE-GROUP main effect is reduced to almost zero when REGION and URBANISM are entered first. This means that the CE groups do not have different retention rates overall, after controlling for REGION and URBANISM. There are REGION x CE-GROUP and URBANISM x CE-GROUP interactions, however, and both tables are marked primarily by unexpectedly high rates for one of the subgroups of the other-CE group—for the Pacific region in one case, and for large districts in the second.

The main conclusion to be drawn from this analysis is that REGION and URBANISM effects are considerably larger than the CE-GROUP effects. Because REGION and URBANISM are variables from the sampling frame, we would expect students within categories of these classifying variables to be relatively homogeneous on many variables, and they do seem to be more homogeneous with respect to retention rates. However, there are also indications that students in the three CE-GROUP categories are unevenly distributed over the REGION and URBANISM categories are unevenly distributed over the REGION and URBANISM categories.

A fairly substantial three-way interaction remains after fitting all the two-way interaction terms, with $\chi^2(16) = 81.26$, $p < .001$. Table 16-16 presents the three-way table of

Table 16-16

Retention Rates and Cell Frequencies for a REGION x URBANISM x CE-GROUP Subtable*

	URBANISM									
	Small			Medium			Large			
REGION	Title I	Other CE	Non-CE	Title I	Other CE	Non-CE	Title I	Other CE	Non-CE	Total
NE	74	(74)	80	63	(53)	61	48	58	60	64
	596	120	2,102	507	232	1,838	819	844	4,139	11,197
SE	(71)	(64)	77	59	—	64	56	—	56	65
	244	159	1,113	261	11	1,081	445	9	940	4,262
MW	(77)	—	(71)	—	—	76	—	—	59	65
	188	3	362	41	54	800	63	85	1,894	3,489
C	65	—	66	57	(53)	54	63	(49)	54	58
	436	80	2,407	556	159	2,151	519	282	3,222	9,811
P	(58)	(63)	64	(48)	(42)	48	51	54	37	50
	156	110	1,009	194	235	1,868	933	1,338	1,206	7,048
Total	70	64	72	58	51	58	53	54	55	
	1,620	472	6,994	1,559	690	7,739	2,777	2,557	11,400	

*Rates omitted for cells with $N < 100$ and enclosed in parentheses for cells with $100 < N < 400$

retention rates and cell frequencies. This table has a feature not encountered in earlier tables, namely, that some of the cell frequencies are too small for reliable estimates of the retention rates. We have omitted rates for cells having frequencies of 100 or below, and rates for cells having frequencies in the 100-400 range are enclosed in parentheses, as a reminder that these rates have relatively large standard errors (from .025 to about .05). Judging from the cell *N*s, other-CE programs affect very few students outside of large districts in the Northeast and Pacific regions. Almost 60 percent of other-CE students are in these two cells. Title I students are much more widely distributed. The three-way interaction for predicting rates may owe more to sampling fluctuations than do any substantive influence. Both the highest and lowest rates are for Non-CE students, with the highest rate in the Northeast-small-non-CE cell (80 percent), and the lowest rate in the Pacific-large-non-CE cell (37 percent).

To help clarify the REGION x URBANISM x CE-GROUP results, analyses were undertaken in which the variables ACHIEVEMENT and ETHNICITY were added to these three variables. These further analyses provided more detail, but scant clarification. Adding more variables seems to make it more difficult rather than less difficult to achieve acceptable fits to the data. Small cell sizes, together with sampling variability and the unreliability of the classifying variables, may be adding noise faster than information.

Table 16-17

Retention Rates and Cell Frequencies for a REGION x URBANISM x ETHNICITY Subtable*

	URBANISM									
	Small			Medium			Large			
REGION	Black	Hispanic	White/ Other	Black	Hispanic	White/ Other	Black	Hispanic	White/ Other	Total
NE	—	—	80	73	54	64	58	45	69	67
	75	28	2,660	488	745	1,260	2,518	471	2,450	10,694
SE	(82)	—	77	67	—	68	68	—	55	68
	232	9	1,229	498	9	744	425	18	881	4,045
MW	—	74	—	—	—	79	54	—	67	68
	19	3	521	5	10	848	739	16	1,159	3,320
C	(76)	—	67	(58)	(62)	57	61	(48)	56	61
	196	39	2,580	259	382	2,040	1,111	109	2,645	9,361
P	—	(67)	69	—	(53)	50	50	59	43	54
	23	394	758	26	240	1,894	8,444	1,184	1,204	6,567
Total	79	67	74	67	[illegible]	61	57	55	59	
	545	472	7,747	1,276	1,386	6,786	5,637	1,798	8,340	

*Rates omitted for cells <100, and enclosed in parentheses for cells with 100 < *N* < 400

For its intrinsic interest, a REGION x URBANISM x EHTNICITY subtable from these further analyses is given in Table 16-17. Again, the rates for cells with *N*s smaller than 100 omitted, and those for cells with *N*s between 100 and 400 are enclosed in parentheses. Examination of the *N*s indicates that except for Hispanic students in the Pacific region, there are very few minority students in small districts, with large districts having the most minority students, especially for blacks. Eighty-three percent of Hispanic students are in either the Northeast or Pacific regions, and another 15 percent are in the Central region, leaving a mere 2 percent in the Southeast and Midwest regions. As indicated earlier, nearly half of the Hispanic students are in the Pacific region, and they have an overall retention rate of .60, which is about average. Hispanic students in the Northeast region have a somewhat lower retention rate (.51). Most Hispanic students in the Pacific region probably have a Mexican or Central American background, while most Hispanic students in the Northeast probably have a Puerto Rican or other Caribbean background. The differences in attrition rate could indicate that analyses involving Hispanic students should distinguish between those in eastern and western parts of the country. Hispanic students in large Northeast districts have one of the lowest retention rates in the table (.45). The lowest rate of all belongs to white/other students in large Pacific districts, however, (.43), which is surprising. If we think of attrition as a risk factor, we would expect minority students to show higher attrition. In three of five comparisons for large districts, white/other students have lower retention than black students.

SUMMARY AND CONCLUSIONS

This study extends the description of the SES representative sample into the second year, with special emphasis on identifying the demographic factors most closely associated with student attrition and on assessing the impact of such factors on longitudinal comparisons among educational programs. A total of 35,808 students, tested at some time during the first year of study, were classified by attendance (attrition) pattern, geographic region, urbanism, race/ethnicity, receipt of free or reduced-price meals, basic-skills achievement, cohort, and selection for compensatory education. The results of these classifications were examined to assess the main and joint effects on retention and then to interpret such effects in terms of the objectives of the longitudinal study.

Attrition is widespread in the nation's schools, which should come as no surprise. In individual schools, pupil retention rates ranged from a low of 28 percent to a high of 87 percent. Overall, only about 60 percent of the students could be tested in the fall of two successive years. About 20 percent of the students tested at Time I departed before the close of the study, and another 20 percent entered the study after Time 1. This way of stating retention rates, as a percentage of the number of students passing through the study, tends to exaggerate attrition to some extent. With these figures, one could rephrase the findings so that there is a relatively stable core consisting or 75 percent of the students, while 25 percent of the students move into and out of the study in any given year. However, changing the base for the rate would not affect conclusions about differential attrition for various subgroups.

We assume that most of the observed attrition occurs when students and their families move into or away from the study schools. We can distinguish three cases: (1) students move because of a permanent change of residence, (2) students are temporarily lost to the study, because of illness, staying with a relative, and so on, (3) students are in the schools, but their data are lost to the study through clerical error, unscorable tests, and other kinds of data collection problems. Case (3) is the only source of attrition which is under our control. Students with a PAP (Present-Absent-Present) pattern are interesting, because while we cannot distinguish between the three cases in the data, case (1) is an unlikely reason for a PAP data pattern—that is, we don't expect many instances where students will make a permanent change of residence twice in a year, moving away from and then back to the same school district. The number of PAP students is small, however, and represents less than 1 percent of the students having achievement test data at any one time. This implies that data collection problems are not a major cause of the observed attrition for the SES.

For purposes of analysis, the simplest models for longitudinal data will be those that use 'pure' longitudinal samples, consisting only of cases with complete data on all variables. This study shows that such a sample will not be the same as any cross-sectional one. There are models for data with missing observations, but they have limitations for the present are models for data with missing observations, but they have limitations for the present study.

Afifi and Elashoff (1966) and Timm (1970) have surveyed the relevant literature. Morrison states the usual caveat. "Of course, in every case it is essential that the causes of the missing data are completely independent of the nature or values of the response variates". (Morrison, 1976, pg 120) But, of course, in non-experimental research this caveat will almost invariably be violated. When the response variate is achievement, it is easy to think of casual relationships with missing data. On the one hand, movement of families can disrupt children's educations, and lead to lower achievement, on the other hand, low achievement may cause some families to transfer their children to different and perhaps better schools.

There is ample evidence that the values for achievement scores are higher for students having stable PPP data patterns than for students in any of the attrition groups. What this shows is that stable students are different from the students who enter or leave, and thus do not represent students generally. That may not be a serious problem, since students entering or leaving the sample are, by definition, not receiving the full benefit of their CE or other educational programs To evaluate a program fairly, we might in principle want to limit the analyses to students receiving the program for at least a full year. A more serious statistical problem is the problem of differential attrition, where different subgroups have different rates of attrition. The most striking instance of differential attrition was obtained for a Title I-non-CE comparison within ethnic subgroups. Other things being equal, and disregarding the opposing effects of regression, white/other students had little differential attrition, Hispanic students had differential attrition which could help Title I look good and black students had differential attrition which could make Title I look bad. Because a smaller number of Hispanic students receive Title I funds, the net effect of these patterns would be to bias a pure longitudinal sample in the direction of making Title I look less effective in comparisons with non-CE programs if regression and sampling effects did not completely counteract the bias. Although it is anyone's guess which exerts a greater influence on the findings, it seems important to incorporate these concerns into the analysis of longitudinal data, even if only by caveat. It is our general opinion that our analyses will be influenced by several miniscule biases, but that they work in opposite directions so that we do not have to be concerned that they will gang-up and seriously distort our findings. The biases are further reduced or eliminated entirely for analyses that use more carefully selected groups than merely non-CE students to serve as the comparisons.

REFERENCES

Adkins, D. C., and Ballif, B. L. *Animal Crackers.* Monterey: CTB/McGraw-Hill, 1975.

Afifi, A. A., and Elashoff, R. M. Missing Values in Multivariate Statistics. I. Review of the Literature. *Journal of the American Statistical Association*, 1966, *61*, 595-604.

Alkin, M. C. , Kosecoff, J., Fitzgibbon, C., and Seligman, R. Evaluation and Decision Making: The Title VII Experience. *CSE Monograph No. 4.* Los Angeles: Center for the Study of Evaluation, University of California, 1974.

American Psychological Association, Inc. *Standards for Educational and Phychological Tests.* Washington, DC: Author, 1974.

Anderson, S., Ball, S., Murphy, R., and Associates. *Encyclopedia of Educational Evaluation.* San Francisco: Josey-Bass, 1975.

Anderson, S. A. Educational Compensation and Evaluation: A Critique. In J. C. Stanley (Ed.), *Compensatory Education for Children Ages 2 to 8.* Baltimore: Johns Hopkins University Press, 1973.

Ayrer, J. E., and McNamara, T. C. Survey Testing on an Out-of-Level Basis. *Journal of Educational Measurement*, 1973, *10*, 79-83.

Baker, R. L. *Measurement Considerations in Instruction Project Development.* Paper presented at Conference on Problems in Objectives-Based Measurement, Center for the Study of Evaluation, University of California, Los Angeles, 1972.

Barker, P., and Pelavin, S. *Concerning Scores and Scale Transformations in Standardized Achievement Tests, Their Accuracy and Dependability for Individual and Aggregation: The Case of MAT 70.* Santa Monica: The RAND Corporation (WN-9161), 1975.

Berry, G. L., and Lopez, C. A. Testing Programs and the Spanish-Speaking Child: Assessment Guidelines for School Counselors. *School Counselor*, 1977, *24 (4)*, 261-269.

Bialer, I. Conceptualization of Success and Failure in Mentally Retarded and Normal Children. *Journal of Personality*, 1961, *29*, 303-320.

Birman, B. F. *Case Studies of Overlap Between Title I and PL94-142 Services for Mildly Handicapped Students.* Menlo Park: SRI International, 1979.

Birren, J. E. Psychophysiology and Speed of Response. *American Psychologist*, 1974, *29*, 808-815.

Bock, R. D., and Yates, G. *MULTIQUAL: Log-Linear Analysis of Nominal or Ordinal Data by the Method of Maximum Likelihood.* Chicago. National Educational Resources, 1973.

Bormuth, J. R. *On the Theory of Achievement Test Items.* Chicago: University of Chicago Press, 1970.

Bormuth, J R. Reading Literacy: Its Definition and Assessment. *Reading Research Quarterly,* 1973, 9, *(1),* 7-66.

Breglio, V. J., Hinckley, R. H., and Beal, R. S. Students' Economic and Educational Status and Selection for Compensatory Education. *Technical Report No. 2 from The Study of the Sustaining Effects of Compensatory Education on Basic Skills.* Santa Ana: DECIMA Research, 1978.

Bridgeman, B Generality of a "Fast" or "Slow" Test-Taking Style Across a Variety of Cognitive Tasks. *Journal of Educational Measurement,* 1980, *17 (3),* 211-217.

Brookover, W. G., Erikson, E. L., and Joiner, L. M. *Self-Concept of Ability and School Achievement.* East Lansing: Human Learning Research Institute, Michigan State University, 1967.

Buros, O. K. (Ed.). *The Mental Measurements Yearbook.* Highland Park: Gryphon Press, 1965, 1972.

Campbell, D. T., and Erlebacher, A. How Regression Artifacts in Quasi-Experimental Evaluation Can Mistakenly Make Compensatory Education Look Harmful. In J. Hellmuth (Ed.), *Disadvantaged Child, Vol. 3: Compensatory Education: A National Debate.* New York: Brunner/Mazel, 1970.

Campbell, D. T., and Stanley, J. D. *Experimental and Quasi-Experimental Designs for Research.* Chicago: Rand McNally, 1966.

Carver, R. P. Two Dimensions of Tests: Psychometric and Edumetric. *American Psychologist,* 1972, *July,* 512-518.

Cleary, T. A., and Hilton, T. L. An Investigation of Item Bias. *Educational and Psychological Measurement,* 1968, *28,* 61-75.

Cohen, J. *Statistical Power Analysis for the Behavioral Sciences.* New York: Academic Press, 1969.

Coleman, J. S., Campbell, E. Q., Hobson, C. J., McPartland, J., Mood, A. M., Weinfield, F. D., and York, R. L. *Equality of Educational Opportunity.* Washington, DC: U.S. Government Printing Office, 1966. Catalog Number FS5.238:38001.

Coopersmith, S *The Antecedents of Self-Esteem*. San Francisco. W. H. Freeman and Company, 1967.

Coulson, J. E., Ozenne, D. G., Van Gelder, N. C., Inuzuka, C., Bradford, C., and Doherty, W. J. *The First Year of the Emergency School Aid Act (ESAA) Implementation*. Santa Monica. System Development Corporation, 1975. (Number TM 5236/008/000)

Crandall, V. C , Katkovsky, W., and Crandall, V. J. Children's Beliefs in Their Own Control of Reinforcements in Intellectual-Academic Achievement Situations. *Child Development*, 1965, *36*, 91-109.

Cronbach, L. J. *Essentials of Psychological Testing (3rd ed.)*. New York: Harper, 1970.

Cronbach, L. J. Test Validation. In L. Thorndike (Ed.), *Educational Measurement (2nd ed.)*. Washington, DC: American Council on Education, 1971.

Cronbach, L. J , and Suppes, P. (Eds.). *Disciplined Inquiry for Education*. National Academy of Education, 1969.

CTB/McGraw-Hill. *The Comprehensive Test of Basic Skills, Form S*. Monterey: Author, 1974.

Davis, F. B., and Diamond, J. J. The Preparation of Criterion-Referenced Tests. *CSE Monograph No. 3*. Los Angeles. Center for the Study of Evaluation, University of California, 1974.

Ebel, R. L *Evaluation and Educational Objectives. Behavioral and Otherwise*. Paper presented at the Convention of the American Psychological Association, Honolulu, 1972.

Educational Testing Service. *A Descriptive and Analytic Study of Compensatory Reading Programs: Final Report*. Princeton: Author; 1975.

Epstein, J. L., and McPartland, J. M. *The Effects of Open School Organization on Student Outcomes*. Baltimore: The Johns Hopkins University Center for Social Organization of Schools, 1975. (Report No. 194)

Evans, F. R., and Reilly, R. R. A Study of Speededness as a Source of Test Bias. *Journal of Educational Measurement*, 1972, *9*, 123-131.

Evans, F. R., and Reilly, R. R. A Study of Test Speededness as a Potential Source of Bias in the Quantitative Score of the Admission Test for Graduate Study in Business. *Research in Higher Education*, 1973, *1*, 173-183.

Fink, A., and Kosecoff, J. *Evaluation Primer*. Beverly Hills: Sage, 1980.

Flaugher, R. L., and Pike, L. W. *Reactions to a Very Difficult Test by an Inner-City School Population: A Test and Item Analysis.* Princeton. Educational Testing Service, 1970.

Glaser, R. Instructional Technology and the Measurement of Learning Outcomes: Some Questions. *American Psychologist,* 1963, *18,* 519-521.

Glaser, R., and Nitko, A. Measurement in Learning and Instruction. In R. L. Thorndike (Ed.), *Educational Measurement (2nd ed.).* Washington, DC. American Council on Education, 1971, 652-670.

Guilford, J P *Fundamental Statistics in Psychology and Education.* New York. McGraw-Hill, 1965.

Harman, D. Illiteracy An Overview. *Harvard Educational Review,* 1970, *40,* (2), 226-243.

Harris, C *Comments on Problems of Objectives-Based Measurement.* Paper presented at the annual meeting, American Educational Research Association, New Orleans, 1973

Harris, L, and Associates, Inc. *Survival Literacy Study.* New York. National Reading Council, 1970.

Harris, L, and Associates, Inc *The 1971 National Reading Difficulty Index. A Study of Functional Reading Ability in the United States.* New York. National Reading Council, 1971.

Harris, M. L, and Stewart, D. M. *Application of Classical Strategies to Criterion-Referenced Test Construction.* Paper presented at the annual meeting, American Educational Research Association, New York, 1971.

Hemenway, J. A., Wang, M., Kenoyer, C. E., Hoepfner, R., Bear, M. B., and Smith, G. The Measures and Variables in the Sustaining Effects Study. *Technical Report No. 9 from the Study of the Sustaining Effects of Compensatory Education on Basic Skills.* Santa Monica: System Development Corporation, 1978.

Hively, W *Introduction to Domain-Referenced Achievement Testing.* Symposium presentation, American Educational Research Association, Minneapolis, 1970.

Hively, W, Maxwell, G., Rabehl, G., Sension, D., and Lundin, S. Domain Referenced Curriculum Evaluation A Technical Handbook and A Case Study From the MINNEMAST Project. *CSE Monograph No. 1.* Los Angeles. Center for the Study of Evaluation, University of California, 1973.

Hoepfner, R., Bastone, M., Ogilvie, V. N., Hunter, R., Sparta, S., Grothe, C. R., Shani, E., Hufano, L., Goldstein, E., Williams, R. S., and Smith, K. O. *CSE Elementary School Test Evaluations.* Los Angeles: Center for the Study of Evaluation, University of California, 1976.

Hoepfner, R., Conniff, W., Jr., Petrosko, J., Watkins, J., Erlich, O., Todaro, R., and Hoyt, M. *CSE Secondary School Test Evaluation: Grades 7 and 8.* Los Angeles: Center for the Study of Evaluation, University of California, 1974.

Hoepfner, R., Stern, C., and Nummedal, S. G. *CSE-ECRC Preschool/Kindergarten Test Evaluations.* Los Angeles: Center for the Study of Evaluation and the Early Childhood Research Center, University of California, 1971.

Hoepfner, R., Strickland, G., Stangel, G., Jansen, P., and Patalino, M. *CSE Elementary School Test Evaluations.* Los Angeles: Center for the Study of Evaluation, University of California, 1970.

Hoepfner, R., Zagorski, H., and Wellisch, J. The Sample for the Sustaining Effects Study and Projections of Its Characteristics to the National Population. *Technical Report No. 1 from the Study of the Sustaining Effects of Compensatory Education on Basic Skills.* Santa Monica: System Development Corporation, 1977.

Hofstadter, R. *Anti-Intellectualism in American Life.* New York: Vintage, 1963.

Horn, J. L. *Concepts of Intellect in Relation to Learning and Adult Development.* Paper presented at annual meeting of the American Educational Research Association, San Francisco, 1979.

Horst, D. P., Tallmadge, G. K., and Wood, C. T. *Measuring Achievement Gains in Educational Projects (Report UR-243).* Bethesda: RMC Research Corporation, 1975.

Horst, D. P., Tallmadge, G. K., and Wood, C. T. *A Practical Guide to Measuring Project Impact on Student Achievement.* Los Altos: RMC Research Corporation, 1975.

Keesling, J. W. *Identification of Differing Intended Outcomes and Their Implications for Evaluation.* Paper presented at the annual meeting of the American Educational Research Association, Washington, DC, 1975.

Kenoyer, C. E., Cooper, D. M., Saxton, D. E., and Hoepfner, R. The Effects of Discontinuing Compensatory-Education Services. *Technical Report No. 11 from the Study of the Sustaining Effects of Compensatory Education on Basic Skills.* Santa Monica: System Development Corporation, 1980.

Khan, S. B. The Relative Magnitude of Speed and Power in SCAT. *Journal of Educational Measurement,* 1968, *5,* 327-329.

Kish, L. *Survey Sampling.* New York: Wiley, 1965.

Klein, S. P. Evaluating Tests in Terms of the Information They Provide. *Evaluation Comment,* 1970, *2(2),* 1-6.

Klein, S. P. *An Evaluation of New Mexico's Educational Priorities.* Paper presented at the Western Psychological Association, Portland, 1972.

Klein, S. P., Fenstermacher, G., and Alkin, M. The Center's Changing Evaluation Model. *Evaluation Comment,* 1971, *2(4).*

Klein, S. P., and Kosecoff, J. B. Issues and Procedures in the Development of Criterion-Referenced Tests. *ERIC/TM Report 26.* Princeton: ERIC Clearinghouse on Tests, Measurement, and Evaluation, 1973.

Levine, R. S., and Angoff, W. H. *The Effects of Practice and Growth in Scores on the Scholastic Aptitude Test.* Princeton: Educational Testing Service, 1958.

Lichtman, M. The Development and Validation of R/EAL, an Instrument to Assess Functional Literacy. *Journal of Reading Behavior,* 1974, *6 (2),* 167-182.

Long, J. V., Schaffran, J. A., and Kellogg, T. M. Effects of Out-of-Level Survey Testing on Reading Achievement Scores of Title I ESEA Students. *Journal of Educational Measurement,* 1977, *14,* 203-213.

Mager, R. F. *Preparing Instructional Objectives.* San Francisco: Fearon, 1962.

McLaughlin, D. H., Gilmartin, K. J., and Rossi, R. J. *Controversies in the Evaluation of Compensatory Education.* Palo Alto: American Institute for Research, 1977.

Miller, T. W., and Weiss, D. J. Effects of Time Limits on Test-Taking Behavior. *Research Report 76-2 from the Psychometric Methods Program.* Minneapolis: University of Minnesota, 1976.

Morrison, D.F. *Multivariate Statistical Methods (2nd ed.).* New York: McGraw-Hill, 1976.

Murphy, R. T. *Adult Functional Reading Study, Final Report.* Princeton: Educational Testing Service, 1973.

Murphy, R. T. *Adult Functional Reading Study, Supplement.* Princeton: Educational Testing Service, 1975.

Nafziger, D. H., Thompson, R. B., Hiscor, M. D., and Owen, T. R. *Tests of Functional Adult Literacy: An Evaluation of Currently Available Instruments. Final Report.* Portland: Northwest Regional Educational Laboratory, 1975.

National Assessment of Educational Progress. *National Assessment of Educational Progress, Reading and Literature: General Information Yearbook. Report 02-GIY.* Washington, DC: U.S. Government Printing Office, 1971.

National Assessment of Educational Progress. *National Assessment of Educational Progress, Reading, Released Exercises. Report 02-R-20.* Washington, DC: U.S. Government Printing Office, 1971.

National Institute of Education. *Demonstration Studies of Funds Allocation Within Districts,* Washington, DC: Author, 1977a.

National Institute of Education. *Title I Funds Allocation: The Current Formula.* Washington, DC: Author, 1977b.

Nitko, A. J. *Problems in the Development of Criterion-Referenced Tests.* Paper presented at the annual meeting, American Educational Research Association, New Orleans, 1973.

Novick, M. R., and Lewis, C. Prescribing Test Length for Criterion-Referenced Measurement. *CSE Monograph No. 3.* Los Angeles: Center for the Study of Evaluation, University of California, 1974.

Nowicki, S., and Strickland, B. R. A Locus of Control Scale for Children. *Journal of Consulting and Clinical Psychology,* 1973, *40,* 148-154.

Oosterhof, A. C. Similarity of Various Item Discrimination Indices. *Journal of Educational Measurement,* 1976, *13, 2,* 145-150.

Ozenne, D. G., Van Gelder, N. C., and Cohen, A. J. *Emergency School Aid Act (ESAA) National Evaluation: Achievement Test Restandardization.* Santa Monica: System Development Corporation, 1974.

Piers, E., and Harris, D. *The Piers-Harris Children's Self-Concept Scale.* Nashville: Counselor Recordings and Tests, 1969.

Plake, B. S., and Hoover, H. D. The Comparability of Equal Raw Scores Obtained for In-Level and Out-of-Level Testing: One Source of the Discrepancy Between In-Level and Out-of-Level Grade-Equivalent Scores. *Journal of Educational Measurement,* 1979, *16,* 271-278.

Popham, W. J. *Educational Evaluation.* Englewood Cliffs: Prentice Hall, 1975.

Popham, W. J., and Husek, T. R. Implications of Criterion-Referenced Measurement. *Journal of Educational Measurement,* 1969, *6(1),* 1-9.

Psychological Corporation. *Survey of School Attitudes.* New York: Author, 1975.

Rindler, S. E. Pitfalls in Assessment Test Speededness. *Journal of Educational Measurement,* 1979, *16,* 261-270.

Roshal, S. M., Frieze, I., and Wood, J. T. A Multitrait-Multimethod Validation of Measures of Student Attitudes Toward School, Toward Learning, and Toward Technology in Sixth Graders. *Educational and Psychological Measurement*, 1971, *31*, 999-1006.

Scriven, M. The Methodology of Evaluation. Perspectives of Curriculum Evaluation. *American Educational Research Association Monograph Series on Curriculum Evaluation, No. 1*. Chicago: Rand McNally, 1967.

Skager, R. *Generating Criterion-Referenced Tests from Objectives-Based Assessment Systems: Unsolved Problems in Test Development, Assembly, and Interpretation*. Paper presented at the annual meeting, American Educational Research Association, New Orleans, 1973.

Skager, R. *Critical Differentiating Characteristics for Tests of Educational Achievement*. Paper presented at the annual meeting, American Educational Research Association, Washington, DC, 1975.

Stanford Research Institute. *Longitudinal Evaluation of the National Follow-Through Program, 1970-1971*. Menlo Park: Author, 1970.

Stricht, T. G., Caylor, J. S., Kern, R. P., and Fox, L. E. Project REALISTIC: Determination of Adult Functional Literacy Skill Levels. *Reading Research Quarterly*, 1972, *7 (3)*, 424-465.

Timm, N. H. The Estimation of Variance-Covariance and Correlation Matrices from Incomplete Data. *Psychometrika*, 1970, *35*, 417-437.

Wang, M., Hoepfner, R., Zagorski, H., Hemenway, J. A., Brown, D. S., and Bear, M. B. The Nature and Recipients of Compensatory Education. *Technical Report No. 5 from the Study of the Sustaining Effects of Compensatory Education on Basic Skills*. Santa Monica: System Development Corporation, 1978.

Wasson, J. B. *The Influence of Time Limits on Reading Comprehension Tests Among Fifth-Grade Pupils of Differing Abilities*. Unpublished Ph.D. Dissertation: University of Minnesota, 1969.

Williams, M. H. Does Grouping Affect Motivation? *Elementary School Journal*, 1972, 130-137.

Wilson, H. A. *A Humanistic Approach to Criterion-Referenced Testing*. Paper presented at the annual meeting, American Educational Research Association, New Orleans, 1973.

Wilson, H. A. A Judgmental Approach to Criterion-Referenced Testing. *CSE Monograph No. 3*. Los Angeles: Center for the Study of Evaluation, University of California, 1974.

Winer, B. J. *Statistical Principles in Experimental Design.* New York: McGraw-Hill, 1971.

Yates, A. J. Level, Speed, and Personality Factors in the Intellectual Performance of Young Children. *British Journal of Educational Psychology,* 1966, *36 (3),* 312-316.

Yoshida, R. K. Out-of-Level Testing of Special Education Students With a Standardized Achievement Battery. *Journal of Educational Measurement,* 1976, *13,* 215-221.

Zweig, R., and Associates. *Personal communication to Drs. Fink and Kosecoff,* March 15, 1973.

☆ U.S. GOVERNMENT PRINTING OFFICE: 1981-785-109/2762