DOCUMENT RESUME

ED 142 569 88 TM 006 367

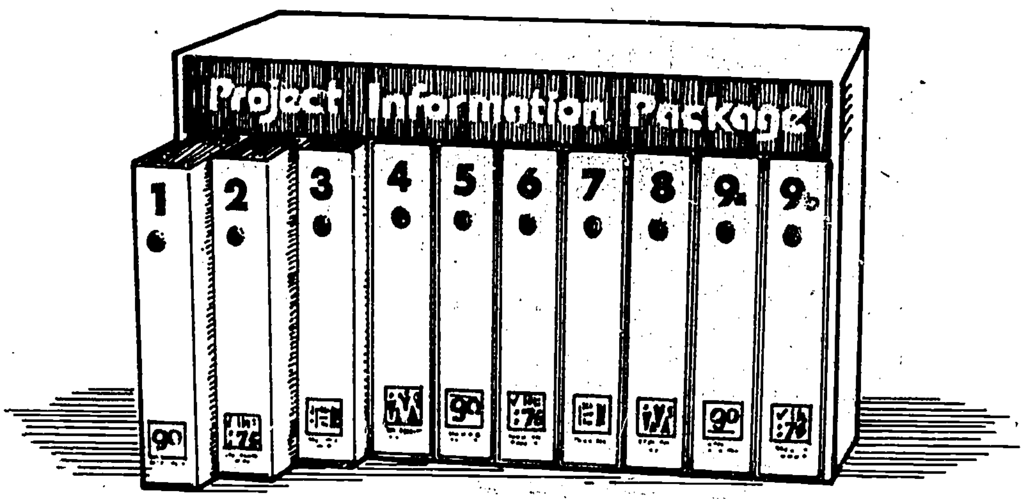
AUTHOR	Norwood, Charles R.
TITLE	Evaluation of the Field Test of Project Information Packages: Volume 2--Technical Report.
INSTITUTION	RMC Research Corp., Los Altos, Calif.; Stanford Research Inst., Menlo Park, Calif.
SPONS AGENCY	Office of Education (DHEW), Washington, D.C.
REPORT NO	SRI-URU-3556
PUB DATE	[77]
CONTRACT	OEC-0-74-9256
NOTE	419p.; For related documents, see TM 006 366, ED 122 373-375, and 460; Tables may be marginally legible due to small type
EDRS PRICE	MF-$0.83 HC-$22.09 Plus Postage.
DESCRIPTORS	Achievement Gains; Achievement Tests; Curriculum Evaluation; Data Analysis; *Demonstration Projects; *Diffusion; Elementary Secondary Education; Evaluation Criteria; Evaluation Methods; Field Studies; Goodness of Fit; Instructional Materials; *Norm Referenced Tests; Program Effectiveness; Remedial Programs; *Statistical Analysis; *Summative Evaluation; Teacher Attitudes; Test Validity
IDENTIFIERS	Equipercentile Growth Assumption; Metropolitan Achievement Tests; *Packaging; *Project Information Packages

ABSTRACT

Project Information Packages (PIPs), kits containing directions for the replication of exemplary remedial programs, were field tested in a number of school districts. Based on curriculum analysis and site visits PIPs were found to induce projects which were adequate copies of what was packaged; however, what was packaged was not sufficient to implement the same curriculum across sites. Particular emphasis was given to the statistical procedures used in comparing the extent to which the PIPs were implemented in the test sites and in determining gains on the Metropolitan Achievement Tests (MAT). The MAT and the associated norm-referenced analysis were closely examined. The assumption that PIPs, if fully implemented in appropriate school settings, would cause an increase in MAT scores was found to be false. This relationship was also not detectable at all grade levels. Analysis revealed that teacher responsiveness, rather than PIP implementation, was more often associated with gains in test scores. Suggestions for future evaluation methods and the manual of procedures for PIP testing are included. (GDC)

EVALUATION OF THE FIELD TEST OF PROJECT INFORMATION PACKAGES:

Volume II Technical Report



- Project Catch-Up
- Project Conquest
- High Intensity Tutoring
- Intensive Reading Instructional Teams
- Programed Tutorial Reading
- Project R-3

Prepared for:
U.S. Office of Education
Department of Health, Education and Welfare
Washington, D.C. 20202

STANFORD RESEARCH INSTITUTE
Menlo Park, California 94025 · U.S.A.

EVALUATION OF THE FIELD TEST OF PROJECT INFORMATION PACKAGES:

Volume II Technical Report

By: CHARLES R. NORWOOD

Prepared for:

U.S. OFFICE OF EDUCATION
DEPARTMENT OF HEALTH, EDUCATION AND WELFARE
WASHINGTON, D.C. 202020

CONTRACT OEC-0-74-9256

SRI Project URU-3556

PREFACE

In my opinion, this technical report is somewhat different from the usual Office of Education evaluation product.

The differences are not so much due to differences in the questions we asked and the difficulties we faced. The differences are due to our idea of what constitutes a relevant evaluation of an educational product that is supposed to raise norm-referenced, standardized achievement test scores.

Enough evidence had accumulated to make us doubt the reasonableness of the assumption that available standardized achievement tests are sensitive to what most compensatory reading teachers are doing in class. Thus, we sought to document facts that would make it plausible that the treatment received by students would teach them what to answer on the test selected as the outcome measure. First, this entailed finding out what was actually done (not merely what was supposed to be done); and second, it entailed an analysis of the connection between what was done and what would teach the items.

Most evaluation products address these two points in an abstract way referring to item domains and teaching objectives, instead of specific items and activities. We have attempted to do a more detailed analysis. The reader will judge how successful we were.

This report also differs somewhat from many Office of Education evaluation products in its technical approach. Where possible, we have not calculated statistical hypothesis tests which are meaningless in this type of evaluation. We tried to create successful inductive arguments based on reported facts, not based on a nonexistent sampling scheme. We also attempted to directly measure our variables for formal analysis. This avoided the necessity of making post-hoc scales for some important variables. For example, we had our site visit staff assess implementation based on their experience; no further scaling was done. Our regression equations were aimed at modeling the bivariate distribution of pretest and posttest scores, as opposed to the distribution of posttest scores given pretest scores. The reader will judge how successful these techinques were.

I attempted to include in this report enough information about our procedures and analyses so that the reader could understand the bases for our conclusions, draw his own conclusions, or carry out the activities himself in another study. The majority of readers interested in the PIP field test and its outcomes are not concerned with the details of the steps we took in our evaluation. These readers are referred to Volume I, the Summary Report.

The original six PIPs, developed by RMC and tested in the field during the school years 1974-1976 are the subject of the evaluation described in this report. These original PIPs are no longer in use; they were thoroughly revised by our subcontractor, RMC, on the basis of first-year evaluation findings.

This report owes its quality to the efforts of the project staff. Our evidence on whether the PIP projects should influence outcome scores was analyzed by Arlene B. Tennenbaum and Christine Miller, assisted by Christine Padilla. Many technical problems associated with our analysis were investigated by David Kaskowitz.

The high validity of our field work is due to the conscientious work of our field visit staff: Dorothy Booth, Jay Cross, Cassandra Duarte, Phillip Giesen, Georgia Gillis, and Margaret Needels. They actively participated in all phases of the project.

Encoding of data and preparation of raw data files was supervised by Bert Laurence and Bill Lambert. The creation of analysis files and execution of analyses were supervised by George Black. George Byrd, Jerry Kauffman, Pat McCall, John Rollin, and Roy Sutton all helped program analyses. Quality control procedures for machine-readable data were supervised by Elizabeth Milrod.

The labor of report writing was shared by Trudy Nio (Chapter 1), David Kaskowitz (Chapter 3), Georgia Gillis (Chapter 4), Marian Stearns, and Christine Miller (Chapter 5), and myself. Elizabeth Milrod, assisted by Nancy Craig, who was project secretary, and Christine Padilla coordinated our efforts.

Marian Stearns was project director and author of Volume I of this report. Ann Bezdek was our project officer.

Charles R. Norwood

SUMMARY

EVALUATION OF THE FIELD TEST OF PROJECT INFORMATION PACKAGES (PIPs)

Purpose of the Study

In its continuing search for successful means by which to disseminate exemplary education projects, the U.S. Office of Education supported the development of six Project Information Packages (PIPs) in 1973. The PIPs were designed to provide "how to" information and instructions to facilitate the installation and implementation of exemplary compensatory reading and mathematics projects in new school districts with a minimum of technical assistance. The projects selected for packaging passed criteria of effectiveness with respect to reading and mathematics skills, as well as meeting other criteria. They were reviewed and passed as exemplary projects by the Education Division Joint Dissemination Review Panel. Four of the projects were for elementary school students and two were for middle or secondary school students. Five of the six projects, originally developed by local education agencies (LEAs) with ESEA Title I funds, were "pull-out" projects. These projects supplemented the regular school reading or mathematics programs rather than replacing them; they served a special target group of students, and required additional space and teaching staff. The sixth project, for junior high school students, originally developed by a local education agency with other federal funds, was not a pull-out project but, rather, served all students in a specified grade and required regular classroom teachers to make changes in their instructional methods.

The central principle assumed in developing the PIPs (the "replication" principle) was that, if the antecedent conditions of the effective instructional project could be established in a new site, the project would be reproduced in the new site, and would again prove effective in terms of student achievement gains. In addition, two other assumptions were made. First, it was assumed necessary to match the setting of the replicating site with the setting of the original, successful site. This was to be accomplished by providing information to potential project adopter sites about the original projects and their settings in an Analysis and Selection Kit (ASK). Districts interested in replicating a packaged project would use the ASK in selecting an appropriate project that matched local conditions. Second, it was assumed that project management was the key to replicating the original conditions. Given this assumption, the PIPs highlighted the importance of a dynamic, experienced project director who had responsibility for orienting others in the school district, hiring staff and providing for inservice training and mobilizing the resources necessary to establish the antecedent conditions for effective instruction. Information provided in the PIPs was management-oriented to help project directors and teachers set up conditions for the original instructional program to be recreated; the packages specified requirements for space, qualified staff, materials and equipment, student selection, scheduling, record keeping, and the like. The PIPs differed in the amount of information they provided about the actual instructional program. All but one did not describe in detail the teaching/learning episode, classroom interaction, or sequences of events within the instructional program. Nor did the PIPs describe the uses of each of the recommended curriculum materials, since it was assumed that the appropriate events would follow, given the appropriate mix of resources and the specified teaching staff.

In fiscal year 1975, PIPs were tried out in a number of sites across the country. Under Section 306 of ESEA Title III, grants were awarded to 19 sites for the purpose of implementing one of the exemplary projects via a PIP at each site. To assess the feasibility of replicating successful projects via packages as a way of improving reading and mathematics skills of disadvantaged children, a contract was awarded to Stanford Research Institute in June 1974. The two-year study was to examine the implementation of the packaged projects in the tryout sites and to focus on the following questions:

- Are local education agencies motivated to adopt a packaged project?
- Can exemplary projects be implemented in new sites via the PIPs? Where implementation problems are due to faults in the packages, can reasonable modifications be recommended?
- What functions and in what amount is technical assistance required? If considerable technical assistance is required, can the packages be made more autonomous?
- Are the projects, implemented via the PIPs, effective in improving student achievement?
- What are the effects of the projects on student attitudes? Are the projects acceptable to the local education agency, to teachers, parents, and the community?
- What is the cost of implementing the projects?

The first-year study objectives were to examine the project adoption and installation processes by documenting the discrepancies in each of the 19 sites from RMC expectations and PIP specifications, determining the usefulness of the PIPs for guiding project implementation, and determining the soundness of the principles and assumptions upon which the PIPs were developed. Also of concern during the initial year were the identification of implementation problems encountered by the tryout sites and recommendations about how the packages might be revised in light of the problems identified. RMC was the subcontractor responsible for the latter formative evaluation tasks.

The second-year study objectives were to investigate the impact of the projects on student achievement, to explain differences from expected outcomes, and to explore the participating school districts' intentions for continuing the projects in school year 1976-77 when Title III, Section 306 money was no longer available. RMC created a new ASK and revised the PIPs based on results of the first year study.

The Study Approach

The central concept of the ASK/PIP dissemination strategy is that of replication: In the right community, the PIP with little or no technical assistance, plus a good faith effort on the part of qualified staff, will replicate the successes of the original project.

To test this concept the first year was devoted to assessing the degree of project installation, and the effectiveness of the principles used to select and package the exemplary projects to be reproduced. Therefore, the first year's major evaluation strategy was to compare projects with PIP specifications.

During the first year, five visits were made to each of the 19 field-test sites to observe the project, to interview project and nonproject personnel involved in installation, and to conduct pre- and post-student testing. Although the first year of the evaluation focused on project installation rather than impact on participating students, both standardized achievement tests and attitude surveys were administered to a minimal sample of participating students to get a sense of likely effects and to make sure that the implementation year was not disruptive in terms of student achievement and attitudes. Observations of the projects during the site visits were used to determine the degree of implementation of the specified elements in the instructional environment. Interviews were conducted with administrative and instructional staff members to learn about the installation process, to determine the causes of implementation problems, project modifications, and deviations from PIP specifications to determine acceptability of the packaged projects and the like. Informal contacts were also made with parents of children who were participants to determine their reaction to the project.

In addition to the site visits, contact report forms were used throughout the year by project staff at the original and tryout project sites and by government and evaluation staff to report telephone conversations, visits, and other contacts in which assistance or clarification was requested, offered, or obtained. Finally, instructional staff questionnaires and administrative staff questionnaires were administered to assess staff attitudes toward the PIP and the local project, and resource/cost questionnaires were administered to determine the resources and associated costs of project installation.

The focus of the second year study was on project effectiveness as measured by achievement test gains.

To ensure that the instructional programs described in the PIPs were well understood, a conference was held in Washington, D.C. At this conference PIP project directors and the project directors at the original sites were brought together. RMC staff members who developed the PIPs were also present. This meeting provided the PIP project staff with direct confirmation, or disconfirmation, of the practices which were intended to be communicated by the PIPs.

After this conference, our evaluation activities in the field were essentially similar to those of the first year, except that in the second year we tested all children, not just a sample as in the first year, and we collected lesson plans and information on project curricula from teachers in the site visit sample.

In the second year we also found it necessary to add some special studies to our evaluation activities.

Results of the First Year Study

The results of our first year study, reported in Stearns (1976), were generally favorable. Projects were implemented to the degree of specificity required by the PIPs, and this was done with relatively minor technical assistance. Student attitudes were not adversely impacted by project participation, and most project staff were enthusiastic.

However, there was little evidence that the PIP projects raised reading scores on the Metropolitan Achievement Test (1970) to the extent it was expected the original projects would have. We could see no reason why this result would not reoccur in the second year. Consequently we conducted some special studies of the analytic foundations of the PIP criterion of success.

Results of the Second Year Study

Special Studies

The special studies examined several assumptions implicit in our concept of what would be evidence for PIP success. Our work may be conveniently divided into two areas. The first is an examination of the properties of the MAT as applied to the PIP evaluation (a test of the replication principle). The second was our examination of the properties of our statistical procedures as applied to the MAT.

The PIP replication principle is rooted in widely held beliefs about the worth and properties of reading comprehension tests of the type represented by the MAT. One of these beliefs is that, except for statistical features, most nationally normed tests are essentially the same. For example, OE has funded the Anchor Test Study in which the MAT and seven other tests were treated as equivalent tests of reading comprehension. Therefore, the replication principle does not assert that the PIP projects will raise test scores on only the tests which the originating projects used. It is usually suggested that such projects would not be worth disseminating, if their effects were so specific.

These beliefs have not been unchallenged. [See Wargo and Green (in press) for papers on both sides of this controversy.] We examined the issue of how specific the effects of the PIP projects would be, based on the premise that in projects where the correct PIP instructional style is used, learning gains would follow from a reasonable length of exposure to the PIP curriculum. We also assumed that if a reading comprehension test is relevant to the PIP curriculum these gains would be detectable.

We therefore examined the PIPs to identify the instructional style that was specified. We found that very little information was provided on the projects' philosophy of learning and instructional practices for teaching children. It is fair to say that, overall, the packages described project installation primarily from a Project Director's point of view; they did not describe instruction from the point of view of a teacher who had never implemented the project before.

That the PIP did not presume to teach how to teach was no oversight on the part of the package designers. The PIPs attempted to establish the preconditions for obtaining achievement gains by focusing on placement of projects in communities with appropriate resources (through the ASK), by focusing on project and classroom management, and by hiring staff who had qualifications similar to those at the original site.

Therefore we determined to analyze the curriculum materials which were used in the projects for their relevance to MAT Reading. We found that projects using a given PIP did not use the same teaching materials, that the materials used were not very different from the regular classroom materials, and that project materials generally were not especially relevant to reading comprehension tested by the MAT. (At the first and second grades, content of the curriculum materials and the MAT Reading subtest corresponded somewhat more closely.) In summary, it would be surprising if the projects, however well implemented, showed remarkable gains on MAT Reading. And since the PIP curriculum was not very different from the regular curriculum, what gains were observed might not be due to the PIP project. A brief examination of the tests used to validate the originating projects showed that the original tests were generally more suited to the curricula.

The investigation of the properties of our statistical procedure as applied to the MAT yielded some interesting formal results, as well as a better understanding of the implications of accepting the MAT norms as longitudinally valid. We concluded the assumption that the MAT norms are longitudinally valid should not be uncritically accepted. Assuming that the MAT norms are longitudinally valid has implications about the growth of Reading achievement which have not yet been independently verified. For example, as measured by the MAT, Reading in children at the 90th percentile in the second grade is growing almost six times faster than is Reading in eighth grade children at the same percentile; Reading in second grade children at the 10th percentile is growing only 1.5 times faster. If this is true it is certainly an interesting fact for developmental psychologists to explain. On the other hand, this interaction of learning rate, grade, and percentile may just be an artifact of the MAT scaling process.

Artifacts may indeed be present. In the fourth grade, where we did not change levels of the test, only one site out of ten gained in the percentile of its mean between fall and spring. At the third grade, eight sites of nine showed gains in percentile, as did all seven sites in the fifth grade. Simulations we performed confirmed these "grade effects" were not necessarily effects of the PIP instructional processes.

Based on our special studies we concluded that the PIP replication principle was not sound: there was little reason to expect that projects which correctly rendered the PIPs would dramatically increase MAT-type reading scores.

Achievement Test Results

Analysis of our MAT achievement test results confirmed the expectation that the Reading scores would not dramatically change.

Norm referenced analyses showed that Reading scores were not (educationally) significantly greater than expected. The projects did generally maintain at least the growth predicted from the norm tables. However, since the growth predicted from the norm tables is generally an underestimate of expected growth, given that the pre- and posttests are not perfectly reliable, achieving more than the expectation calculated from the norm tables is not as impressive as it at first seems.

Bivariate regression equations were fit to the pre- and posttest scores, using dummy variables to encode the effect of well-implemented teachers. We did not find that being in a well-implemented class was systematically related to gains larger than the gains in poorly- or moderately well-implemented classes.

CONTENTS

EXHIBITS

ILLUSTRATIONS

TABLES

1 INTRODUCTION

The evaluation of Project Information Packages (PIPs) during the second year of their field test is described in this volume. We present the rationale, the data collection and analysis procedures, and our findings with respect to instructional implementation and student achievement outcomes. In this introductory section, we describe the history of the packaging concept and each of the six PIPs created for the U.S. Office of Education (USOE) by RMC Research Corporation. In addition, we describe the field test of the PIPs and the organizations involved.

1.1 Background of the PIPs

As part of its desire to spread innovative practices from federally funded projects and to improve federal program performance, USOE supported the development of six PIPs. These packages were conceived as an alternative to methods of dissemination that require demonstration, on-site training, or other hands-on technical assistance.

The basic concept underlying the development of the PIPs was replication. The Office of Planning, Budgeting, and Evaluation (OPBE) within USOE conceived of a package of instructions that would enable a school district to replace an ineffective educational program with an educational program that had proved effective in another district. If such a project could be established at a new site as specified in the instructions, the assumption was that the project would prove effective again in terms of producing equivalent student achievement gains. Presumably, the equivalent gains would occur because the essential program features or set of learning conditions had been replicated by the new site.

Another consideration in the packaging concept was that the package be a complete and self-contained collection of information and instructions assembled and structured so as to enable educators at a new site to select a program and implement it. The package was to contain sufficient do-it-yourself information to reproduce the program without assistance.

In the spring of 1973, RMC Research Corporation was commissioned to create packages that provided "how to" information and instructions to facilitate the implementation in new school districts of selected compensatory reading and mathematics programs. With efforts geared to finding eight successful reading and math programs, RMC developed criteria for selecting exemplary programs for packaging and embarked on an elaborate search and review procedure. Over 2000 projects were suggested or recommended for consideration; after an initial screening, over 100 were subjected to a thorough review.

Three screening criteria were used by RMC to select programs for packaging: (1) effectiveness, (2) cost, and (3) availability and replicability.

Effectiveness--RMC reviewed the results of a local evaluation of each project to determine the validity of the data and the claims of effectiveness.

> The effectiveness criterion had two separate aspects, statistical and educational significance. Both were cast in terms of (a) the pre- to post-test gains of project participants and (b) the gains which would have been expected had they not received the special treatment. The educational significance criterion which was agreed upon specified that observed gains had to exceed expected gains by at least one-third standard deviation with respect to the national norms. The statistical significance criterion specified that the difference between observed and expected gains had to reach or exceed the five percent confidence level using a one-tailed statistical test (Tallmadge, 1974, p. iii).*

Because there were almost no cases in which a control group had been used in the local evaluations, RMC decided that "expected gains" for students in compensatory programs would mean that the group remained in the same percentile on both a standardized pretest and a posttest. Any group that exceeded these expected gains by at least a one-third standard deviation (with reference to the norm distribution of the particular test) met RMC's criterion for educationally significant gains. For one project,

---

*A list of references is appended to this report.

Programed Tutorial Reading (PTR), RMC used the regression-discontinuity model to determine effectiveness.*

RMC reported that selection of programs that were successful enough to be recommended turned out to be extremely difficult. "Not one of the several hundred project evaluations which were examined provided acceptable evidence regarding project impact" (Tallmadge, 1974, p. iv). Even where adequate data had been collected, analysis and reporting practices forced inferences to be drawn with caution. To meet the effectiveness criterion, RMC decided that evidence of statistical and educational significance had to be found in two instances, for example, at two different grade levels or for two different years. Under these conditions, only six projects met the effectiveness criterion and the other criteria.

Cost--Although attempts were made to find effective programs that were also inexpensive--to make it more likely that programs would be exportable to many sites--a recurring cost of $475 per student was eventually established as the upper limit for projects selected, with the additional provision that start-up costs not exceed $1000 per student.

Availability and Replicability--These areas were more judgmental and were based on the fact that the packages were likely to be disseminated--at least on a trial basis--under the auspices of the federal government. Projects were rejected on the criterion of availability if directors and staff were unwilling to cooperate in helping the packaging effort or if the project was no longer operating and could not be visited for validation. Projects were also rejected if they were not operating in public schools or if their selection would amount to a USOE endorsement of a single publisher's or manufacturer's commercial product(s). Projects were rejected on the criterion of replicability if they required resources not generally available in typical school districts. This consideration included highly unusual personnel, equipment, or environments. Projects rejected on replicability grounds included a university-operated elementary school and a project requiring major architectural modifications to the school building (Tallmadge, 1974, p. iii). Since USOE felt that PIPs

---

*RMC published a practical guide to project evaluation on the basis of their experiences in searching for exemplary programs. SRI assumed that the procedures described in this booklet by Horst, Tallmadge, and Wood (1975) were those RMC endorsed. These procedures form the basis of the norm-referenced analysis we used to evaluate student achievement in the PIP field test.

might eventually be disseminated under Title I auspices, RMC excluded from their search those programs that were not aimed at reading and mathematics achievement for disadvantaged children in kindergarten through twelfth grade.

Although RMC had wished to find eight projects, only six satisfied all the criteria. RMC prepared the program descriptions and the evidence of effectiveness for USOE's Dissemination Review Panel,* and all six projects were approved as exemplary projects. They were:

Project Catch-Up	Newport-Mesa Unified School District Newport Beach, California
Project Conquest	School District 189 East St. Louis, Illinois
High Intensity Tutoring (HIT)	School District of the City of Highland Park Highland Park, Michigan
Intensive Reading Instructional Teams (IRIT)	Hartford Public Schools Hartford, Connecticut
Programed Tutorial Reading (PTR)	Davis County School District Farmington, Utah
Project R-3	San Jose Unified School District San Jose, California

1.2 Nature of the PIPs

After identifying the six PIP programs, RMC developed the packages that contained the "how to" information. PIPs were boxes (18 X 14 X 12 inches) with ten upright drawers (containing nine components) and one large flat drawer at the top (see Figure 1-1). Inside the drawers were brochures, manuals, filmstrips, cassettes, catalogues, charts, and the like.

---

*This panel is now a joint activity of USOE and the National Institute of Education (NIE) and is called the Joint Dissemination Review Panel (JDRP).

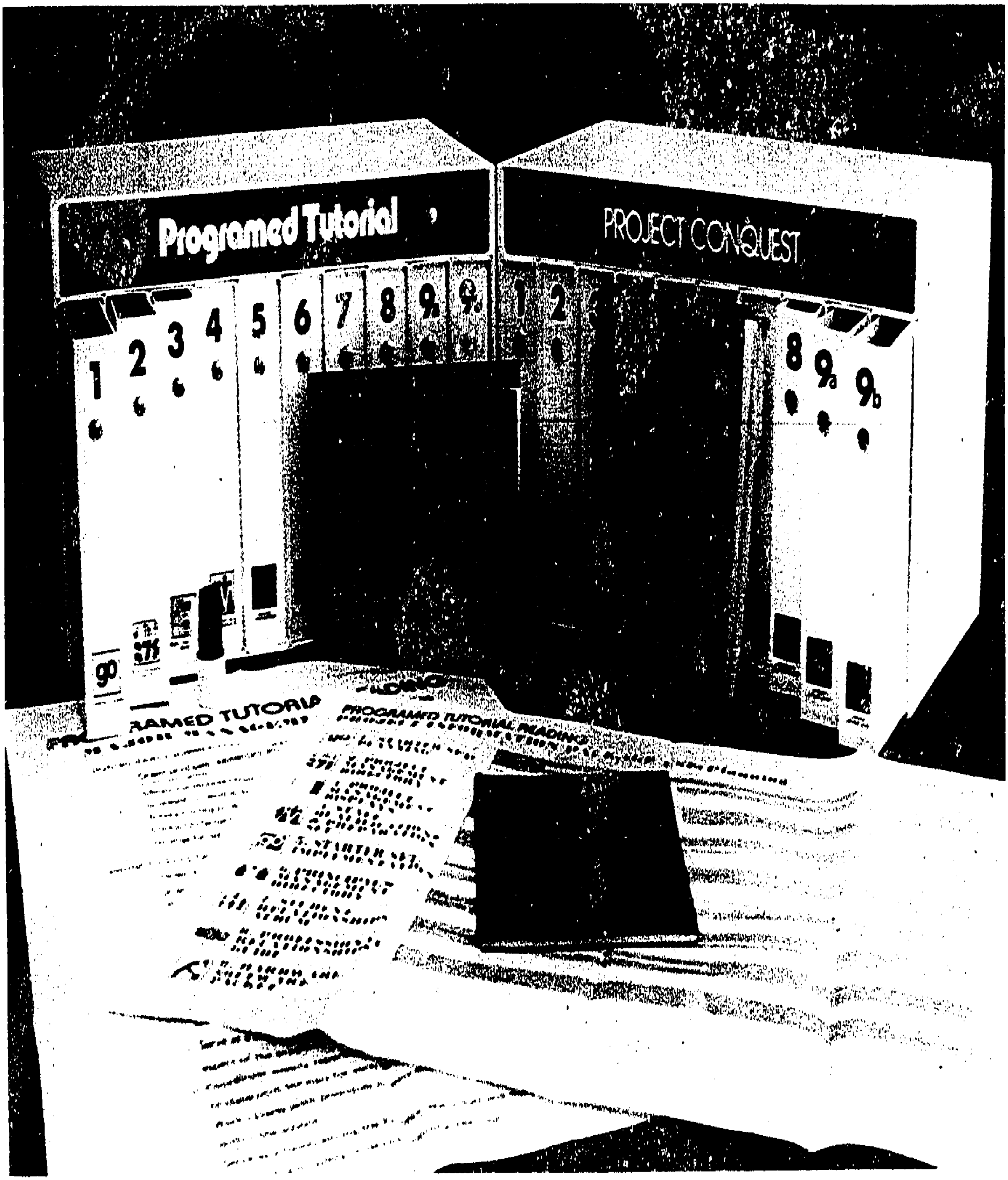


FIGURE 1 PROGRAMED TUTORIAL READING AND PROJECT CONQUEST PIPS

1.2.1 PIP Components

> The first four components were concerned with planning and other preimplementation tasks required to equip, staff, and set up a new project in a school district. They were primarily intended to assist the director in installing the adopted project. The remaining five components were more specifically concerned with project implementation and were directly related to the day-to-day operation of the project after it was installed (Tallmadge, 1974, p. v).

The components are described below according to their functions and contents. Table 1-1, adapted from an RMC report (Piestrup, 1974), summarizes the PIP components and their purposes.

Starter Set: Planning--This component provided the new project director with multimedia materials to use in presenting the project and in gaining support of parents, teachers, principals, and school boards. The starter set described the features of the project for a general audience with some elaboration on these features so that the project director could conduct subsequent briefings and answer questions on the project. This component included one-page handouts, illustrated brochures, a filmstrip, and a tape cassette.

Project Management Directory--This manual brought together in calendar form the list of management activities for the project. Key tasks in installing and implementing the project were blocked by weeks so the director could fill in the actual event, such as a meeting with the principal, on the day it would occur. The directory summarized the tasks for each month in a checklist to emphasize the importance of accomplishing the tasks as close as possible to the scheduled time. It also provided budget-updating summaries for each month for the director to fill in.

Project Management Displays--The displays were designed to attract attention to the existence of the project and to elicit questions or interest concerning it. The displays provided a time schedule overview and summarized the key tasks to be performed in developing the model project and the components to be used in performing the tasks.

Staff Qualifications and Preparation Set--This set provided project directors with information on hiring personnel. It contained a staff

Table 1-1

SUMMARY OF PIP COMPONENTS

Component	Personnel/Purpose	Contents
Installation (start-up) stage 1. Starter set: planning	For project director: Provides orientation to project Provides public relations materials on project and package Provides introduction to package (and director's role in project) For school boards, principals, regular school staff, potential project staff, parents: Provides information to elicit support	Filmstrip with cassette tape Handout brochure Project director's orientation booklet (with cassette tape for some projects) Viewfoils
2. Project management directory	For project management personnel (project director, assistant director, principal, as applicable): Provides detailed guidelines and support materials needed to plan and implement (operate) the project	Project management calendar Supplementary sections on the major management tasks
3. Project management displays	For project director and visitors: Provides time schedule overview Summarizes component use Displays component use and time schedule to visitors	Major management tasks chart PIP use displays
4. Staff qualifications and preparation set	For director: Provides personnel selection guidelines For staff: Provides self-evaluation and orientation materials	Staff qualifications checklist In-service training topics
Implementation stage 5. Starter set: implementation	For project classroom personnel: Helps in starting each type of activity (testing, teaching, other) including setting up environment for the first time	Implementation starter booklet Original art file
6. Classroom management directory	For project classroom personnel: Provides guidelines for classroom procedures Provides samples of materials needed for administration of the project, record forms, letters and notices, and the like	Teaching staff guide (e.g., a calendar and support materials)
7. Student relationships album	For project staff interacting with children: Describes the project environment, from the child's viewpoint, that staff is expected to create (e.g., how he should perceive staff; what learning climate he should experience--high pressure, self-directed, and so on) Distinguishes roles of different staff members in creating environment Describes desired student responses (e.g., confident, competent, happy, eager) and gives specific instances	Album
8. Professional relationships guide	For all project staff (plus principal): Defines roles in relation to all school staff (project and nonproject) with whom each project member interacts Provides descriptions of potential staff conflicts and suggested resolutions	Guide booklet
9. Hardware software packet	For project director and teaching staff: Aids in selection and ordering of commercial hardware/software Provides description, source, and features of core and supplementary items For teaching staff: Describes experience of original project staff plus modification (if any)	Factsheets and manufacturers' brochures for core items Supplementary materials list with publishers' addresses, available factsheets, and brochures

Source: Piestrup (1974).

qualifications checklist, information on training topics, and a suggested agenda for training project teachers and aides.

Starter Set: Implementation--This component provided information for the early weeks of instruction in the new project and included materials for decorating the classroom, a detailed calendar for the first two weeks of school, and descriptions of how to begin each new activity (testing, teaching, and the like). This starter set enabled new staff to create an attractive environment appropriate for the region where the project was to be implemented.

Classroom Management Directory--This directory, designed to correspond to the Project Management Directory in format and purpose, contained day-by-day management instructions for the classroom teacher. Calendar entries indicated the sequence of events and reminded teachers to perform key tasks throughout the year. The directory included monthly task summaries in a checklist, budget records, and supplementary sections explaining alternative strategies for accomplishing tasks and anticipating problems and described the practical details of operating an instructional system.

Student Relationships Album--This album described the roles of staff members in relation to students. Role illustrations were presented to convey the "flavor" of the project from a child's viewpoint. The album dealt with the difficult areas of attitude change, self-concept development, fostering achievement orientation, and eliminating prejudice.

Professional Relationships Guide--This guide described problems and conflicts that occurred at the exemplary site and suggested ways of forming good working relationships in the new project. The guide defined the roles and responsibilities of project and nonproject staff in relation to one another.

Hardware/Software Packet--This packet provided information on instructional materials and equipment, some of which were optional. The packet contained publishers' and manufacturers' brochures describing the materials and instructional hardware to be used.

Flat Art File--This file contained reproductions of display materials developed by the original project and instructions for the production of other materials. It enabled the adopting school to make use of project-developed materials designed to enhance motivation, to improve self-concepts, or to facilitate skill development.

In creating the content materials for the components of the PIPs, RMC assumed the following principles:

- It was necessary to match the setting of the replicating site with the setting of the original site. An Analysis and Selection Kit (ASK) was designed to provide local education agency (LEA) staff with information by which to choose the exemplary program best suited to their needs and most likely to be fully implemented in the context of their school district. (See Section 1.2.2 for a description of the ASK.)
- The project director at the new site was the critical element, and management of the installation process was the key to replication of the original program in the new site. Information provided in the PIPs was described to enable project directors to set up the conditions for the effective instructional program to be recreated.
- Project directors were to select instructional staff who were exemplary teachers in the same mold as those in the original project. It was assumed that teachers could not be trained to have certain instructional theories and methods but that they could be selected to fit the project model.
- Minimum information to guide staff training and to guide instructional practices would be sufficient if staff selection were appropriate.
- A list of curriculum materials used at the original site was sufficient to permit replication of the originally effective curriculum and instructional program, since qualified staff would choose appropriately among the materials to serve the individual needs of the students.

1.2.2 Analysis and Selection Kit

Because the introduction of the packages to potential users and the processes within the LEAs for selecting and adopting a program were not encompassed by the PIPs, an Analysis and Selection Kit (ASK) was designed. The prototype kit consisted of the following:

- A brochure for wide dissemination, describing the six exemplary programs and their availability in the form of PIPs. (This kind of brochure had been available for other package dissemination programs such as that sponsored by the Right-to-Read program.)
- Further information, available upon request, about each program. (This information was to enable an LEA to choose the program best suited to its needs.)
- A set of orientation materials (actually the very materials contained in the "Starter Set:Planning" component of the PIP) to enable LEA staff to orient school board, community, instructional staff, and the like to the selected program.

The ASK had barely reached the prototype stage when USOE found an opportunity to test the PIPs in the field.

1.3 Description of the Six Original Programs

The six exemplary programs, as they existed at the time they were examined by RMC, are described below. Table 1-2 summarizes the characteristics of each.

1.3.1 Project Catch-Up

Catch-Up in the Newport-Mesa Unified School District in California provided remedial instruction in reading and mathematics to low-achieving children in a low-socioeconomic suburban area. The program served children from preschool through twelfth grade. (The Catch-Up PIP described only the elementary portion of the program, kindergarten through sixth grade.) The major instructional emphasis of the exemplary program was on diagnosis of learning problems through extensive use of criterion-referenced materials. A special staff of certified teachers diagnosed student problems and prescribed activities and materials for teaching two or three specific skills. The learning experience was provided by certified part-time teachers and instructional aides in an attractive "laboratory" environment. This lab was set up within an available classroom in the school and operated only in the morning when both staff and students were fresh. Students attended the lab 20 to 60 minutes daily, depending on their need for extra reading and mathematics instruction.

It was essential to the philosophy of the program that each teacher be responsible for the achievement gains of no more than 18 students.

Table 1-2

CHARACTERISTICS OF THE SIX EXEMPLARY PIP PROGRAMS

Program	Catch-Up Newport Beach, California	Conquest East St. Louis, Illinois	High Intensity Tutoring (HIT) Highland Park, Michigan
Subject areas	Reading and math	Reading	Reading and math
Students served	Students in grades K-6 who need help in reading or math; average student-teacher ratio, 5:1.	Students in grades 2-6 who are two years or less below grade level, and first grade repeaters; student-teacher ratio, 6:1.	Students in grades 6 or 7 one to five years below grade level tutored by students in grades 7 or 8, at least two years ahead of them on a one-to-one basis.
Nature of program	Students come from classrooms to the Catch-Up lab every morning. The lead teacher, the two other teachers, and the aide work independently with children for 30 minutes. Parent aides assist in instruction.	Students come from classrooms to reading rooms (grades 1-3) or clinics (grades 4-6) for 45 or 50 minutes each day. Heavy diagnostic testing determines areas that the lab clinicians will work on with each student.	One tutor works with one tutee using programmed curriculum for 20 minutes under supervision of a reading or math specialist and two aides. Rewards are earned by tutors as well as tutees.
Program	Intensive Reading Instructional Teams (IRIT) Hartford, Connecticut	Programed Tutorial Reading (PTR) Farmington, Utah	Project R-3 San Jose, California
Subject areas	Reading	Reading	Reading, math, and social studies
Students served	Students in grades 3 and 4 who need help in all areas of language arts; student-teacher ratio, 15:1.	First graders who need help learning to read are tutored on a one-to-one basis.	In the project's first year, all seventh graders in heterogeneous classes (gifted and slow learners); student-teacher ratio, 20:1.
Nature of program	Forty-five students from several schools are brought to one site each morning for 11 weeks for intensive instruction by three language arts specialists in three one-hour sessions.	Students are tutored apart from their regular classrooms for 15 minutes a day by paraprofessional tutors; programmed kits designed for the basal reader used in the classroom are the only curriculum materials used.	Forty-five minute periods of reading, math, and social studies offer an integrated curriculum oriented to gaming/simulation, individualized instruction, and an intensive involvement study trip. A cadre staff (one teacher per subject) moves with students to grades 8 and 9.

1.3.2 Project Conquest

Conquest in East St. Louis, Illinois, used a clinical approach to reading and served students in grades K-6 who were two years or less below grade level. The instructional staff diagnosed each child's reading problems through an intensive 17-step diagnostic procedure and prescribed a structured learning program to be followed by each child. Remedial instruction was provided in 45-minute sessions held four to five days per week. The teachers, called clinicians, received extensive in-service training and supervision in remediation techniques, testing, diagnosis, and related areas. The centers were designated as clinics (for grades 4-6) or reading rooms (for grades K-3) and were separate from the regular classroom, serving several groups of students per day.

The Conquest student experienced three or four activities in the following areas during each session: programmed reading, comprehension, phonics/vocabulary/sight words, and oral or recreational reading. Instruction in at least one of the areas was assisted through the use of a teaching machine.

1.3.3 High Intensity Tutoring (HIT)

HIT in Highland Park, Michigan, was a peer tutoring program that used highly structured materials for sixth, seventh, and eighth graders. Older students tutored younger ones daily in reading or mathematics, or both, using programmed and drill materials. Both tutors and tutees were performing one to three years below grade level on standardized tests. Tutoring was fast-paced and intense. The percentage of correct responses for each tutee in the program was calculated daily to ensure that presentation of new materials was adjusted to the student's rate of learning; the goal was for students to achieve a correct-response rate of 90%-94% each day. Interaction between tutor and tutee was structured to maximize the time that each tutee was engaged in active learning behaviors. Tutors checked each answer as it was made and provided correct answers and reinforcement according to a structured procedure. Rewards were an incentive device for both tutors and tutees. A unique feature of the HIT program was that the tutor also improved academically during the learning process.

There were two HIT centers at each school (a reading center and a mathematics center). Each was staffed by a certified teacher and two paraprofessional aides. To control the student error rate, teachers and aides monitored the tutoring, distributed rewards, and kept detailed records. Teachers administered Sullivan pretests to establish an entry level into the materials for individual students.

1.3.4 Intensive Reading Instructional Teams (IRIT)

The IRIT program in Hartford, Connecticut, was designed to raise the level of language and reading achievement of third and fourth grade pupils who were deficient in the basic skills of language and reading. IRIT employed a team-teaching approach with three specialists--for phonics, individualized reading appreciation, and vocabulary/comprehension. Diagnostic testing identified the special needs of each student. The individualized reading sessions offered reading assignments that enriched the child's background, promoted his written and oral language skills, and instilled pleasure in reading. The vocabulary and comprehension sessions built perceptual and reading skills.

The IRIT program was divided into three 11-week cycles. Forty-five students were selected for each of the three cycles. During the cycle assigned to them, students left their regular classrooms to go to the IRIT classes for three hours each morning, five days a week. The 45 students in the program were heterogeneously divided into three groups of 15 each. During the three-hour morning, each group of 15 students spent one hour on phonics, one on individual reading, and one on vocabulary and comprehension.

1.3.5 Programed Tutorial Reading (PTR)

PTR* in Farmington, Utah, provided tutoring to underachieving first graders in beginning reading as a supplement to conventional classroom teaching. The tutoring materials included the same basal readers that were used in the regular classrooms, along with a comprehension and word analysis book and word list cards.

The teaching strategy employed many of the elements of programmed instruction: frequent and immediate feedback, specified format, and individualized pace. However, whereas programmed instruction had often sought errorless or nearly errorless learning with many cues at first, followed by a fading of cues, the PTR proceeded in the opposite manner with minimal cuing at first, followed by increased prompting until the child could make the correct responses.

First grade students were tutored on a one-to-one basis by carefully trained tutors for 15 minutes each day. The tutors were paraprofessionals who ranged in skill and experience from high school students to teacher aides and community members or parents.

*The PTR program was originally created by Douglas Allson at the University of Indiana.

1.3.6 Project R-3

"R-3" stands for student readiness, subject relevance, and learning reinforcement. The original program in San Jose, California, was a junior high school program for reading, mathematics, and social studies in which the teachers engaged in team planning to create a highly motivating program for the students. All seventh grade students participated in R-3; the program followed these students as they moved on into the eighth grade, and then into the ninth grade. Each student attended one 45-minute period in each of the three subject areas daily. Because R-3 served an entire grade, the group was heterogeneous (i.e., not composed only of low achievers).

Important components of the R-3 program were gaming/simulation activities, learning contracts, individualized instruction, intensive involvement study trips, and parent involvement. Gaming/simulation reinforced skills learned in each subject area. Contracts encouraged the student to set his own goals and to work independently; each student was held responsible for completing his contract. Teachers used an eclectic approach to instruction and to their use of instructional games, simulation, contracts, and intensive involvement. Project staff made home visits to involve the parents in the child's program.

1.3.7 Summary

The six exemplary programs were different from one another. Some were for elementary grades, others for middle-school. Some required non-professionals as instructors; others required highly qualified reading specialists.

Although each program was unique, certain features were similar. For example, two of the six programs, PTR and HIT, were tutorial programs in which instruction was provided on a one-to-one basis. This allowed immediate feedback, a specified format, and an individualized pace. The few curriculum materials used were programmed to maximize the amount of time the tutee was engaged in active learning. Both projects employed supervised students or paraprofessional staff to monitor the tutoring process. Three programs, IRIT, Catch-Up, and Conquest, were laboratory programs. They all used a diagnostic-prescriptive instructional approach in which learning deficiencies were identified and a curriculum prescribed for each student according to his needs. Students in these programs spent more time receiving treatment than did students in the tutoring programs.

All except Project R-3 were "pull-out" programs. That is, they supplemented rather than replaced regular classroom teaching; they required students to leave class and go to another location to participate. In addition, except for R-3, these programs did not require regular classroom teachers to make changes in their methods or behavior other than adjusting schedules so that children could be released to participate in the project.

In R-3, all seventh grade students participated and continued in the program during the eighth and ninth grades. Unlike the other projects, which emphasized reading or reading and math, R-3 included three subject areas: reading, mathematics, and social studies.

1.4 Field Test of the PIPs

In February 1974, while RMC was developing the PIPs, the Title III program office announced the availability of four types of grants under Section 306 and sent application requirements to all school districts. The announcement of Title III grants contained no information about the six PIP programs except that they were reading and mathematics programs for disadvantaged elementary and secondary students. Although RMC had sent OPBE some of the information LEAs would need for selecting a program suited to their needs, they had not yet put information about all of the six PIP programs into a brochure.

After a joint review of applications by Title III program officers, OPBE staff, and consultants, 21 PIP grants were awarded. One school district, which received grants for both the PTR and the HIT PIPs, returned the award after its board of education refused to approve participation in the program. Since two districts had been assigned two PIPs each, the final tally was 19 projects in 17 school districts. Table 1-3 lists the school districts participating in the field test of the PIPs.

The field test of the six PIPs was an opportunity for OPBE to try out the PIPs in school districts and to evaluate the feasibility of the PIPs as a method by which USOE could disseminate programs that would improve the reading and mathematical skills of disadvantaged children. OPBE called for a "comprehensive portrayal and analysis of the process" from which to decide whether to continue, terminate, or modify the six PIPs and the packaging concept.

The principal criterion by which USOE intended to judge the effectiveness of PIPs was that of educationally significant growth in achievement. From USOE's point of view, if the student achievement test outcomes in the

Table 1-3

SCHOOL DISTRICTS PARTICIPATING IN PIP EVALUATION

PIP	Project Location	School District
Catch-Up	Bloomington, Indiana	Monroe County Community Schools
	Brookport, Illinois	Brookport School District
	Galax, Virginia	Galax City Public Schools
	Providence Forge, Virginia	New Kent County Public Schools
	Wayne City, Illinois	District No. 100
Conquest	Benton Harbor, Michigan	Benton Harbor Area Schools
	Cleveland, Ohio	Cleveland Public Schools
	Gloversville, New York	Gloversville Enlarged School District
HIT	Lexington, Mississippi	Holmes County School District
	Olean, New York	City School District of Olean
IRIT	Bloomington, Indiana	Monroe County Community Schools
	Oklahoma City, Oklahoma	Oklahoma City Public Schools
	Schenectady, New York	Schenectady City School District
PTR	Canton, Mississippi	Canton Public Schools
	Dallas, Texas	Dallas Independent School District
R-3	Charlotte, North Carolina	Charlotte-Mecklenburg Schools
	Lake Village, Arkansas	Lakeside Public Schools
	Lorain, Ohio	Lorain City Schools
	Schenectady, New York	Schenectady City School District

field-test sites showed gains equal to those of the original programs, the PIPs could be judged successful in terms of program effectiveness. For effectiveness to be determined, pre- and post-tests of participating students were required. USOE specified the following criteria for selecting test batteries:

> The selected battery is to include subtests, at all grade levels, to assess vocabulary, reading comprehension, word analysis skills, mathematics concepts, reasoning and computation skills. The selected battery (or batteries) must further meet the following criteria:
>
> 1. Provide evidence of subtest reliability and validity.
> 2. Contain interlocked forms for grades kindergarten through 12 (K-12).

3. Contain alternate forms for pre- and post-testing.

4. Be basic-skills oriented (i.e., subtest items should be independent of specific curriculum content).

5. Be of an appropriate difficulty level for the student sample.

6. Include in the norming sample students similar to those in the treatment and comparison groups of the field test.

7. Be easily administered and scored.

8. Require a reasonable amount of administration time (HEW Request for Proposal 74-40, p. 18).

In the test selection task, it seemed appropriate to consider the seven standardized achievement tests included in the Anchor Test Study.* These tests were: the California Achievement Test, the Cooperative Test of Basic Skills, the Iowa Test of Basic Skills, the Metropolitan Achievement Test, the Sequential Tests of Educational Progress, the SRA Achievement Tests, and the Stanford Reading Tests.

Of these instruments, the Metropolitan Achievement Test (MAT) was the only one for which both fall and spring normative data were available. Since it met more of the criteria than did the other tests, it was scheduled for the PIP field-test evaluation. Discussions with personnel at the UCLA Center for the Study of Evaluation revealed that the 1970 edition of the MAT was among the highest rated tests in the CSE Elementary School Test Evaluation report (Hoepfner et al., 1970).

1.5 Organization of the Evaluation

During the past two years many groups were involved in the evaluation study. As mentioned earlier, the Office of Planning, Budgeting, and Evaluation within USOE initiated the packaging concept. The Division of Centers and Supplementary Services in USOE's Bureau of School Systems awarded grants under ESEA Title III, Section 306, to 17 school districts to implement the PIP projects.

Each of the 19 tryout projects had an organizational configuration based on the instructions in its PIP. Each of the six PIPs called for a

---

*The Anchor Test Study (Loret et al., 1974) was the result of USOE's attempt to devise an approach for equating scores from different reading tests.

project director to manage the project, to hire and train the instructional staff, and to orient other district personnel and the public to the PIP program.

USOE awarded a contract to SRI to evaluate the PIPs in their two-year field test. SRI in turn subcontracted with RMC for the formative evaluation and a revision of the six RMC-developed PIPs. During the first year of the study, RMC staff accompanied SRI staff on field visits. During the second year, RMC made revisions based on the findings of the first-year evaluation and continuing input from SRI staff members as the main evaluation progressed. The second-generation PIPs represent attempts to fill the gaps and to resolve the ambiguities identified during the evaluation of the first-generation PIPs.

To foster understanding of their education programs, directors of the exemplary projects provided assistance to both SRI and RMC. The exemplary projects also provided some minimal assistance, upon request, to the PIP field tryout sites.

The SRI evaluation focused on the degree of implementation of the PIP-specified installation activities at each site during the first year and on the effectiveness of the projects in terms of student achievement during the second year. In both years the evaluation plan called for extensive fieldwork through observations, interviews, student testing, and collection of information regarding curriculum materials. Distinct groups performed the functions required for each of these data collection activities. Site visitors were the basic data collection unit. Each of these six professionals was expert in at least one PIP program. Their function was to visit their assigned project sites several times a year to observe the PIP implementation process and to interview the educators on site. It is upon their judgment that our findings about degrees of implementation depended.

Although student achievement was the primary focus of the second-year evaluation, achievement and attitude tests were given to students during both years. SRI used its own personnel to administer the pre- and post-tests in fall and spring. At sites having a large number of students to be tested, SRI administrators hired and trained qualified personnel at the test site to give and monitor the tests according to SRI procedures.

Because SRI personnel are stationed in Menlo Park, California, site assistants were hired at each of the project sites to help the SRI evaluation staff. These local site assistants were hired by the site visitor upon recommendations made by the local project director. The principal duties of the site assistant were to set up the logistics of the site visits and to perform the clerical tasks associated with the testing of project students (e.g., scheduling testing, identifying students to be tested, labeling test booklets, and shipping test materials back to SRI). In addition, when needed, the site assistant provided SRI with copies of local documents, such as student lesson plans and lists of curriculum materials used by the PIP instructional staff.

1.6 Evaluation Issues

The central evaluation issue for the second year report is the validity of the replication principle. This concept is that the PIPs would induce projects which raise achievement test scores to the same degree as the original projects, if the original project and PIP project were given the same norm-referenced achievement test.

This idea of replication is rooted in some widely held beliefs about the worth and equivalence of some nationally normed achievement tests. First it is believed that raising scores on these tests is desirable and, secondly, it is believed that, at least among the tests used in the Anchor Test Study (Loret et al. 1974), most achievement tests are equivalent. Thus, the replication principle does not assert that PIP effects are test specific. Indeed, many policy makers would argue that projects with such specific effects would not be worth disseminating.

Our first year results, which showed definitely unspectacular test gains, led us to question the validity of the replication principle. However, we could not definitely assert that we would not see the expected results for this report.

Consequently, we decided to execute three types of analyses. The first analysis used the concepts underlying the replication principle. These results are reported in Section 3.7. However, based on our first year's experience we knew that the analysis techniques used in Section 3.7 were flawed. The full extent to which they were flawed and the importance of the flaws were not clear. Therefore we undertook a second type of analysis, called a special study. The special studies examined, both analytically and empirically, some of the assumptions underlying the replication principle. The analytic results are discussed in Section 2.3 and the empirical results are in Section 3.8. We call the third type of analysis we did a curriculum-referenced analysis.

The curriculum-referenced analysis is described in Sections 4, 5, and 6. This analysis was based on the premise that test scores go up in response to a well-implemented, appropriate presentation of a curriculum which is relevant to the test items. In our interpretation of this concept, the curriculum must, in the judgment of a competent specialist, appear relevant to the test items. We assumed that the PIP-specified teaching style had already been shown to be relevant. Presumably if PIPs worked, the well-implemented teacher would show better gains than poorly-implemented ones, if the MAT was equally suited to the PIP-specified (and used) curriculum. The extent to which we were able to show this effect is discussed in Section 6.

To summarize, the main evaluation issues were:

- The validity of the replication principle and the associated norm-referenced analyses.
- The results of the norm-referenced analyses: Did the replication principle predict our test results?
- The type and impact of formal errors in the norm-referenced analysis.
- The results of the curriculum-referenced analysis.

The curriculum-referenced analysis involved several further issues:

- What was the PIP-specified instructional style?
- To what degree was this style implemented by teachers?
- What was the PIP-specified curriculum?
- To what extent was this curriculum used?
- Is the Metropolitan Achievement Test sensitive to the curriculum which was specified and used?
- Did teachers who were well-implemented show better gains on the MAT (given the age, race, and sex distributions of their classes) than poorly- or so-so-implemented teachers?

The next section discusses some of the analytic foundations of the replication principle and the associated norm-referenced analyses.

2 STRATEGY OF THE SECOND-YEAR EVALUATION

2.1 Introduction

The Analysis and Selection Kits (ASKs) and the Project Information Packages (PIPs) under scrutiny in the SRI field evaluation were thought of as examples of "strong packaging." That is, they were prescriptive instructions for selecting and implementing specific educational projects. Each set of instructions was intended to be, by itself, sufficient for implementing a copy of a previously successful project, provided of course that the project personnel were willing to make a good-faith effort to follow the ASK-PIP instructions.

The fundamental motivating principle behind PIP packaging was faith in the "black-box" concept of scientific investigation. This familiar research model was adopted from engineering problems in which a closed system, the black box, is assessed by measuring its inputs and outputs. The events inside the system are not necessarily important for predicting outcomes; it is necessary only to relate observable inputs to observable outputs.

The design of the PIPs applied this input-process-output concept at several levels of detail. At the grossest level, when the ASK instructions were followed, the PIPs were to be put into a socioeconomic context that matched the context of the originating site. The PIP itself was to supply the "process" of education. Educationally significantly improved achievement test socres were to be the expected output. At a finer level of detail, the PIP design told the would-be implementer how to hire qualified staff, how to order materials, and how to manage relations between project and nonproject staff and between project staff and students' parents. These events were the input to the black box that was thought of as representing the educational process. The output of the process was to be the already-mentioned educationally significant gain in children's test scores.

The ASK-PIP combination, then, was designed to replicate an originating project by duplicating the project's obvious inputs: socioeconomic context, staff, community relations, and materials. The educational process in the originating site was, by and large, thought of as a black box. As a consequence, the PIPs relied on projects having hired teachers who already knew these details. The PIPs relied on having the correct input; the the output was assumed to follow.

This volume of the report on the results of the PIP field trials will focus on how the black-box approach to strong packaging has fared. In particular, we will examine the claim that the PIPs would produce projects that would cause students to achieve the educationally significant growth that the original projects did.

Our measure of growth was based on the 1970 Metropolitan Achievement Test (MAT). The obvious first step for detecting educationally significant growth on this or any other test was to define "educationally significant growth" so that it could be identified as an output of the originating project; it would then be possible to detect the same event on the output of the replicating project.

As discussed in Volume One, the originating projects were selected in part on the basis of their standardized achievement test outcomes. Consequently, the process for selecting the originating sites defined the sense of "educationally significant growth" that was appropriate to this evaluation.

The selection process, as described in the Dissemination and Review Panel literature, called for the analysts to prefer projects that could demonstrate that students in the project had an average growth of one-third of the norm group standard deviation over that growth expected without the project. Projects Catch-Up, HIT, IRIT, and R-3 were chosen in this way. PTR was chosen on the basis of treatment-control group achievement test differences.

All projects were originally evaluated using different standardized achievement tests, except for Conquest, which was selected without reference to any normed test.

The black-box model predicted that the replicating projects would achieve the same criterion (relative to educationally significant growth) that the originating sites achieved. This criterion however, was not a function of the project only; it was also a function of the measurement methods associated with the several normed tests used in the selection of originating projects. This is unfortunate because the theory and the application of nationally normed achievement tests are not very satisfactory. It became necessary to perform the evaluation while dealing with technical problems not necessarily relevant to the central issue: Are PIPs examples of a dissemination strategy that is worthy of continued federal support?

To remain focused on this central issue, we designed our evaluation in two segments. The first segment follows the spirit of the black-box model, accepting the measurement theory which it assumes when applied to the PIP projects. The second segment reanalyzes the test data from a point of view that is not as dependent on the measurement theory. When data are examined from two points of view, the conclusions they imply may differ. The PIPs may be successful if judged by the black-box approach and not successful if the associated measurement theory is rejected. We, as evaluators, have an obligation to state which analysis we credit. We feel that the best analysis rejects the measurement theory implicit in the black-box outcomes. In section 2.2, we describe our two analysis designs and explain our preference.

2.2 Justification for Norm-Referenced Analysis

Insofar as achievement test scores are concerned, the rational selection of projects suitable for strong packaging was accomplished by assuming what is equivalent to "generic true-score test theory," as discussed by Lord and Novick (1968, p. 164). (Nontest criteria for the selection of projects are described in Volume One.)

The theory of generic true scores treats observed test scores as if they were based on the following sampling scheme:

- Randomly select n persons from the infinite population of humans.
- Randomly select N tests from the infinite population of achievement tests; administer them K times to each person.
- Assume that the $K \times N$ test scores on these tests are independent for a given individual.

We may then write the following standard random effects analysis of variance model for an observed test score, Y_j. Let $E_j Y_{jk}$ be the expectation of Y with respect to the population indexed by j; then the model is:

$$Y_{gak} = u + A_a + P_g + AP_{ag} + \epsilon_{gak} \quad , \qquad (2\text{-}1)$$

where

g indexes tests	g = 1, G
a indexes individuals	a = 1, A
k indexes replications	k = 1, K

$$A_a = T_a - u$$
$$T_a = E_g E_k Y_{gak}$$
$$u = E_g E_k E_a Y_{gak}$$
$$P_g = D_g - u$$
$$D_g = E_a E_k Y_{gak}$$
$$AP_{ag} = E_k Y_{gak} - u - A_a - P_g.$$

ε_{gak} is the specific error of measurement; A_a is the effect of the ath individual; P_g is the effect of the gth test; AP_{ag} is the interaction of the ath individual and the gth test.

Using this conceptualization, we may talk about the various tests used to assess the outcomes of the originating projects as being parallel achievement test forms.* In practice, of course, the forms are only nominally parallel (see Lord and Novick, 1968, p. 174). Obviously, the choice of a particular test is not important, as long as it is a member of the universe of acceptable tests. As discussed in Section 3, our choice--the MAT--is an acceptable measure of outcomes in the sense of generic test theory.

The user of the model (Eq. 2-1) conceives of estimating the generic true score, T_a, defined as the expectation of Y_{gak} over the hypothetical universe of tests and replications. In this sense, for a given individual, all the sampled achievement test scores in the universe estimate the same thing--not the same specific true score, $E_k Y_{gak}$, but the same generic true score, $E_g E_k Y_{gak}$. It is possible to use the random effects model (Eq. 2-1) to estimate generic true scores and their variances if there is a data set in which persons were given multiple tests. Unfortunately, no such data were available on the set of potential originating projects, so the generic true-score model was not used in the selection process, except to justify treating different normed tests as generically equivalent.

Based on the idea of generic equivalence, if only one test were available from the universe of tests, the best estimate of the generic true score for a group of project children is that group's average of observed scores. Our best method of comparing programs is on these av-

---

*The Anchor Test Study (1974) applies these ideas to the MAT and 7 other tests. The obvious criticism of the Anchor Test Study and the PIP replication principle is that they both ignore differences in item content.

erages. A reasonable plan for selecting projects that increase generic test scores would be to select projects that moved the average observed test scores higher than would be expected without the project.

The two questions that immediately arise are how much more than expected, and how much is expected. The first question was settled, somewhat arbitrarily, by accepting projects that had a growth of one-third of the norm group standard deviation over expected growth. The question of how to define "expected growth" was answered by using a control group average, if one was available. If one was not, the test's norm tables were used to estimate expected growth, as follows:

- Compute the average test score in the fall. Call this F.
- Find the fall percentile corresponding to F. Call the fall percentile F'.
- Find F' in the spring percentile tables. Let S be the corresponding spring score. S is taken as the expected spring average score.

Evidently, there is a "sampling error" involved in F, S, and the observed spring score, O. This error should be taken into account by making justifiable assumptions about the joint distribution of F and O, as estimators of average generic true scores, and then deducing what a reasonable statistical test might be. However, the procedure actually followed was to assume that

$$T_{N-1} = \frac{O - S}{\sqrt{\dfrac{S^2_x + S^2_y - 2r_{xy}S_xS_y}{N - 1}}} \qquad (2\text{-}2)$$

is approximately distributed as student's t with N - 1 degrees of freedom, where

O = observed average spring test score

S = expected average spring test score

S_x = observed pretest standard deviation

S_y = observed posttest standard deviation

r_{xy} = observed pre-post correlation

N = number of children with pre- and post-scores.

Calculation of this statistic is called a "norm-referenced analysis," after Horst, Tallmadge, and Wood (1975).

The parts of the evaluation that are true to the design principles of the PIPs are based on the conceptual measurement associated with Eq. 2-1. The ability of the replicating projects to move generic test scores is assessed by computing the norm-referenced analysis based on the value of standard scores obtained from the 1970 version of the MAT. If the lower limit of the 95% confidence interval, based on the t test (Eq. 2-1) for the difference between observed and expected scores, exceeds one-third the norm group standard deviation, we can say that the criterion has been achieved.

2.3 Special Analytic Studies of the Norm-Referenced Analysis

As stated, we feel the norm-referenced analysis should not be taken as the principal test of the validity of the PIP concept. The measurement framework imposed by the adoption of generic test theory, and the computation of the norm-referenced analysis, should be viewed as devices for expediting the selection of projects for packaging. These devices should be viewed, and were viewed, as rules of thumb, not as serious models for selecting exemplary programs. The next paragraphs provide our justification for these remarks.

The generic true score may well be a useful construct for situations in which tests are actually sampled and in which there are replications of tests and samples of persons who have taken them. Certainly, the corresponding random effects analysis of variance has found useful applications. However, there was nothing in the selection of the PIP originating projects to justify the introduction of generic true scores; there was only "conceptual" sampling. Because generally only one test was associated with each originating project, it is impossible to compare the originating projects on estimates of their generic scores.

As a result, we are forced to acknowledge that the PIP projects were selected on the basis of several specific scores, no matter what the conceptualized sampling may have been. Similarly, we must view the MAT achievement scores as specific scores because we cannot compute any of the parameters associated with the generic scores, since the MAT was the only test given. The results of the norm-referenced procedure are not operationally comparable to the validation of the originating project although, in the sense of generic true-score theory, they are (conceptually) estimating the same thing.

In short, the norm-referenced analysis of the MAT achievement data is not operationally identical to the "educationally significant" output that was found at the originating sites. We do not know whether the originating site would have passed the MAT-based norm-referenced test of educational significance.

Additional reasons for not regarding the MAT norm-referenced analysis as critical stem from the properties of the MAT standardization and from the t test (Eq. 2-2).

2.3.1 Criticisms of the MAT Standardizing Procedure

A norm-referenced procedure is only as valid as the norms on which it is based. The 1970 Metropolitan Achievement Test, published by Harcourt Brace Jovanovich, Inc., was made to 1970 test industry standards. The test is in some ways superior to other tests of that period, principally because the MAT has empirical norms for both fall and spring.*

The test makers provide a table of standard scores that are supposed to enable users of the MAT to test out of level and to measure directly the "amount of achievement" on an interval scale. The publishers state:

> The standard score scale for the Metropolitan Achievement Tests provides two basic conveniences for the test user. The scale makes forms within a battery equivalent and provides a continuous, equal interval system for each test across all material. Once raw scores are converted to standard scores, one need not be concerned in further interpretation with either the battery or the form from which the raw scores came (MAT Guidelines No. 1).

We feel that these claims are too strong for the standardization program that was used. Our principal concern is that norming and scaling were done on a cross-sectional, not a longitudinal, basis. In addition, the members of the MAT battery were equivalenced by the equipercentile method without regard for a student's grade. We feel that equivalencing within grades would have been preferable, especially in the lower grades where growth is rapid.

The standardization procedures that were followed have the potential for failing in at least two ways. First, because the equivalencing of the various members of the battery was not made grade-specific, the articulation of the tests between grades may not be good. Second, the cross-sectional norms may mask the MAT's sensitivity to "cohort effects." If the test is sensitive to such effects, the longitudinal implications of the norms may not apply to either the norm sample children or the PIP children. For example, children in the eighth grade MAT sample may well have had educational experiences in the elementary grades much different

---

*Only fall data were used in the standardization program.

from the experiences of the children in the sample's elementary schools. This would mean that the growth curve implicit in the MAT norms is not what would have been observed had the norm group's second graders, for instance, been followed.

There is evidence consistent with the existence of both poor articulation between members of the MAT battery and the existence of cohort effects in the MAT norms. Pelavin and Barker (1976) examined the articulation between the various members of the MAT test battery by testing disadvantaged children twice within seven days on members of the MAT series. Since no large change in the children's achievement could be expected in a seven-day period, one would anticipate that a given child would get the same standard score both times, even if he took a different test each time. Pelavin and Barker found only weak evidence that disadvantaged children get the same score both times. On this evidence, they conclude that evaluators who are predominantly concerned with educationally disadvantaged students should base their evaluation on something other than standardized test scores.

Further support for their conclusion can be obtained from the MAT raw-score to standard-score convernsion tables. Figure 2-1 shows selected equipercentile standard-score growth curves for the Total Reading subscales of the MAT. The test batteries and subscales are described in more detail in Section 3. In Figure 2-1 the growth curves change abruptly between the first and second and the seventh and eighth grades for Total Reading. Presumably, the "sampling error" in these graphs is small, so we should regard the changes as real.* Is this how achievement scores change, or are the fluctuations cohort effects? We note that whatever the fluctuations are, they are not uniform across percentiles. For example, in the Total Reading curves, there are some interesting dips between the fourth and sixth grades for the curves greater than the 80th percentile; these dips do not appear on the lower equipercentile curves. A curious feature of Total Reading score growth in the seventh grade is that for children of the 50th percentile or less no growth is expected until the summer, at which time the curves fairly shoot up on the standard score scale.

On the whole, "summer growth" for Total Reading is about as large as the "school year growth." This means that the growth during May and early June, plus the growth in September and early October, is nearly equal to the growth in the seven months of instruction between October and the following May. If there is not an abrupt change in the learning rate in May or October, we must conclude that there is significant growth in most MAT-relevant skills when there is no school. Expressed in another

*Whether the changes are real or not, the norm-referenced analysis treats them as real. The variations are assumed to be valid and exactly known.

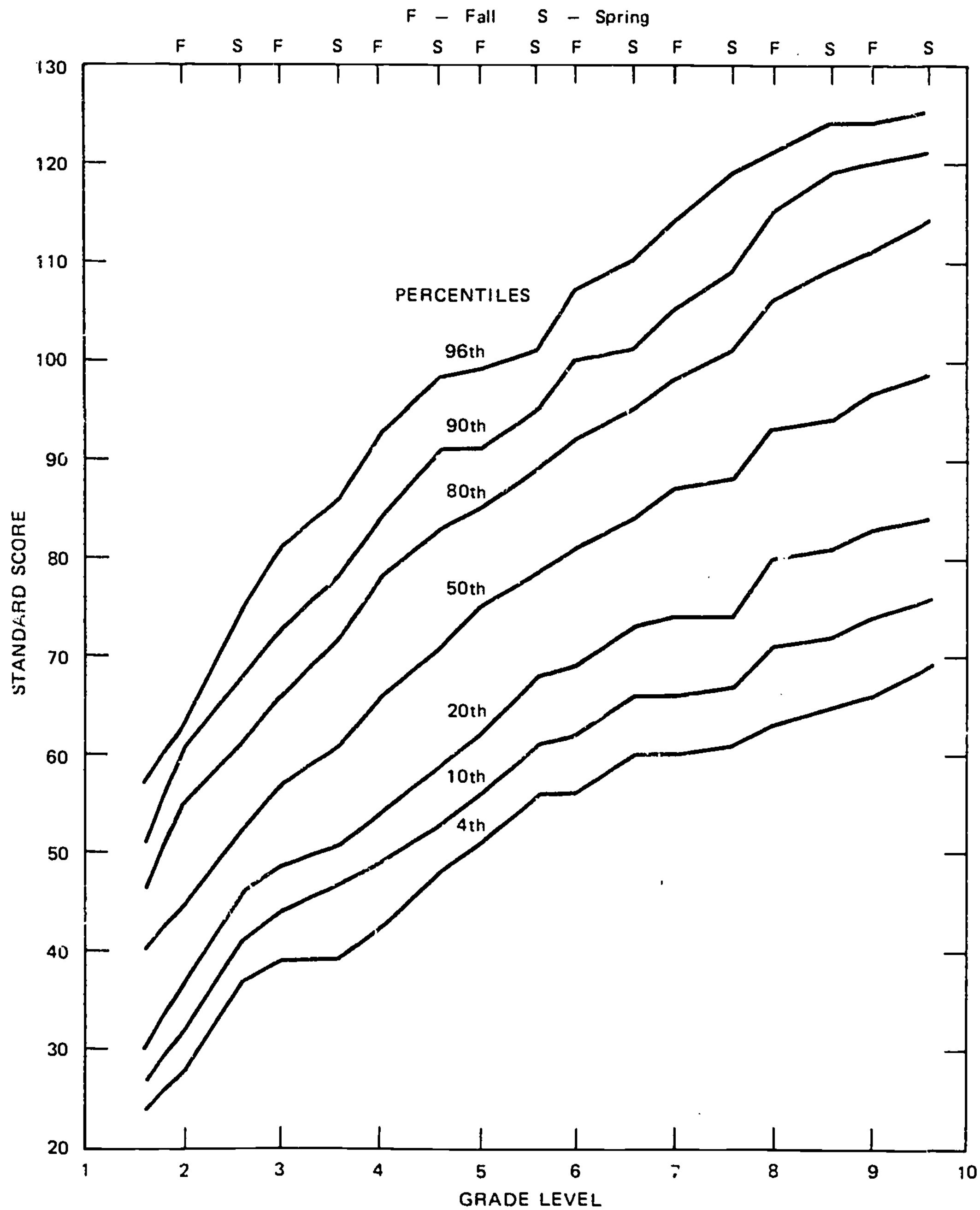


FIGURE 2-1 STANDARD SCORES BY GRADE LEVEL FOR SELECTED PERCENTILE RANKS: MAT TOTAL READING

way, the MAT is measuring skills that grow about as fast out of school as in.

Table 2-1 shows the average of the ratios of spring to fall growth to the following fall to spring growth for Total Reading and Total Math for selected percentiles. The table shows great variation in the average

Table 2-1

AVERAGE OF THE RATIOS OF SPRING-TO-FALL GROWTH
TO THE FOLLOWING FALL-TO-SPRING GROWTH
FOR TOTAL READING AND TOTAL MATH,
FOR SELECTED PERCENTILES

	Number of Ratios with Nonzero Denominators		Average Ratios	
Percentile	Math	Reading	Total Math	Total Reading
1	9	7	0.106	0.250
4	8	7	1.395	0.454
6	7	8	1.519	0.787
10	7	8	1.068	0.988
20	7	7	1.047	1.661
50	7	8	1.242	1.724
80	9	8	1.203	1.025
90	9	8	0.775	1.473
96	9	8	0.434	0.833
99	9	4	0.456	0.872

summer growth as a function of percentile, with the center percentiles being most subject to it. It is particularly interesting that on the average about 70% more growth is found in May, June, and September than in October through April.

Since the MAT norms are cross-sectional, we do not know whether these findings represent facts or artifacts. The Coleman (1966) report and the Jencks (1972) study have both shown that which school is attended does not influence standard reading scores very much. It seems that such findings are built into the norm tables; however, we do not know whether the findings are based on the way MAT skills grow, or on poor test linkages, or on cohort effects.

It is possible to discuss the problems of cohort effects and linkages in the context of the classical test theory that was used to guide the MAT's construction. Within the classical point of view, we could derive models of growth that included cohort effects, item difficulties and the consequent subscale difficulties. This work would form the statistical background for a model that represented "PIP effects." However, we did not do this exercise because our analysis of the content of the MAT items convinced us that the MAT does not represent a good "sample" of the PIP curricula.

Our content analysis, which is discussed in detail in Section 5, also suggests that the "trait" that the MAT measures should not be called "achievement" for programs that are highly individualized. Individualized programs are so narrowly focused that children acquire the specific skills that enable them to accomplish specific tasks. The concept of achievement underlying the MAT is oriented to general skills that may be specialized to increasingly difficult tasks. In the MAT system the degree to which a child has the skill is quantified as proportional to the difficulty of the items he can do. If it is true that the individualized instruction that is supposed to result from PIP implementation causes children to learn specific skills, we would expect that the item difficulties as calculated from PIP children who had the same curricula would be much different from those calculated from the normative data. Unfortunately, as will be seen in Section 5 of this report, the number of children who have had similar curricula is not large enough to allow a convincing test of the issue.

In concluding this discussion of the MAT norms, we note that the utility of standardized tests for evaluations is a well-discussed problem. The basic issues are not empirical ones, but matters of judgment to be discussed in terms of the tester's objectives. Some policies require that educational innovations significantly affect scores relative to a norm group before the programs can be considered useful. In this case, the standardized test is criterial no matter what the objectives of the particular innovations are. We consider such policies rational, if they are clearly and deliberately defined in terms of the test's content and the composition of the test's norm group. Based on our analyses of the correspondence of the PIP curricula and the MAT content, discussed in Section 5, we feel that the use of the MAT for evaluating the PIPs does not meet this condition, except in the nonempirical sense of general test theory.

Before describing our alternative to the norm-reference analysis, we discuss technical difficulties with the "t test" (Eq. 2-2) and with the one-third standard deviation criterion for educational significance.

2.3.2 Criticisms of the Norm-Referenced t Test

The central t distribution is defined as the quotient of a standardized normally distributed random variable and the square root of the quotient of an independently distributed X^2 variable and its degrees of freedom. For the usual application, this would be:

$$t_{n-1} = \frac{\frac{\sqrt{n}\,(\bar{x} - u)}{\sigma}}{\sqrt{\frac{\Sigma(x_i - \bar{x})^2}{(n - 1)\sigma^2}}} = \frac{\sqrt{n}\,(\bar{x} - u)}{\sqrt{\frac{\Sigma(x_i - \bar{x})^2}{n - 1}}}$$

where x_i $i = 1,\ldots, n$ are independently and normally distributed with expectation u and variance σ^2, and where:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \quad .$$

The model for the t test described in Section 2.2 evidently would have x_i being the difference between the observed and expected score for observation i. Formalizing the equal percentile score transformation, the following paragraphs derive a distribution for the difference between the average expected and the average observed spring scores.

We may describe the processes of calculating the equipercentile score as follows. Let

$$z_1 = \ell_1\,(s_1) = \frac{s_1 - u_1}{\sigma_1}$$

$$z_2 = \ell_2\,(s_2) = \frac{s_2 - u_2}{\sigma_2} \quad ,$$

where s_1 is the fall standard score and where u_1 and σ_1 are the expected values and variances of the fall scores. Similarly, ℓ_2 is the corresponding function for the spring.

Assuming the standard scores are normally distributed, the percentiles of s_i, i=1,N are given by:

$$\Phi[\ell_i(s_i)] = \int_{-\infty}^{z_i} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2} du \quad .$$

Since Φ and ℓ_i are invertible, we may write the formula for the equipercentile standard score as:

$$\begin{aligned} s_E &= \ell_2^{-1}\left(\Phi^{-1}\left\{\Phi[\ell_1(s_1)]\right\}\right) \\ &= \sigma_2 \, \Phi^{-1}\left[\Phi(z_1)\right] + u_2 \\ &= \sigma_2 \, z_1 + u_2 \\ &= \frac{\sigma_2}{\sigma_1}\left(s_1 - u_1\right) + u_2 \quad . \end{aligned}$$

For notational simplicity, we may set σ_2/σ_1 to b; then

$$s_E = b\left(s_1 - u_1\right) + u_2 \quad .$$

The result is that the equipercentile score is a linear function of the fall scores.

In fact, the equipercentile method of predicting the spring score is the same as the usual regression method, if it is assumed that fall and spring scores are perfectly correlated. If this correlation, c, is positive and if $b > o$, the equipercentile expectation will be less than the regression expectation for scores less than the fall true mean. For in this case,

$$o < c < 1 \text{ and } cb < b \quad ,$$

but because

$$\left(s_1 - \mu_1\right) < o \quad ,$$

we have

$$cb\left(s_1 - \mu_1\right) + \mu_2 > b\left(s_1 - \mu_1\right) + \mu_2 \quad .$$

Therefore, the equipercentile prediction is an underestimate, from the regression point of view, for scores less than the mean. If $s_1 - \mu_1 > o$, the last inequality reverses. This says that, for fall scores greater than the true fall mean, the equal percentile expectation is too large, from the regression point of view.

This bias will operate to make the PIPs achieve expectation more easily than if the fall-spring correlation in the norms were taken into account.

If we assume that the fall and spring scores are jointly normal with parameters u_1, u_2 σ_1, σ_2, and ρ, where ρ is the fall-spring correlation, the joint distribution of the spring and equipercentile scores is bivariate normal with parameters u_E, u_2, $b\sigma_1$, σ_1, and ρ, where u_E, the expected value of the equipercentile scores, is equal to u_2. If we assume that S_{Ei}, S_{2i}, =1,N are a random sample from the joint distribution of expected and observed spring scores, it can be shown that

$$t_{N-1} = \frac{\overline{S}_2 - \overline{S}_E}{\sqrt{\dfrac{\hat{\sigma}_2^{\,2} + (b\hat{\sigma}_1)^2 - 2\hat{\rho}b\hat{\sigma}_1\hat{\sigma}_2}{N}}} \tag{2-3}$$

is distributed as t on n - 1 degrees of freedom, where $\hat{\sigma}_2^{\,2}$, $\hat{\sigma}_1^{\,2}$, and $\hat{\rho}$ are the usual unbiased estimates of the corresponding parameters, and where

$$\overline{S}_2 = \frac{1}{N}\sum_{i=1}^{N} S_{2i} \text{ and } \overline{S}_E = \frac{1}{N}\sum_{i=1}^{N} S_{Ei} .$$

This formula differs from that given by Horst, Tallmadge, and Wood by the presence of b and by the substitution of N for their N - 1. For moderately large samples, the effect of substituting N for N - 1 is negligible. The effect of b may be seen by the following argument. Suppose $b > 1$, then we have the inequalities:

$$b\hat{\sigma}_1^{\,2} > \hat{\sigma}_1^{\,2}$$

if $b\hat{\sigma}_1^{\,2} - 2\hat{\rho}\hat{\sigma}_1\hat{\sigma}_2 > 0$, then

$$b\left(b\hat{\sigma}_1^{\,2} - 2\hat{\rho}\hat{\sigma}_1\hat{\sigma}_2\right) > \hat{\sigma}_1^{\,2} - 2\hat{\rho}\hat{\sigma}_1\hat{\sigma}_2$$

$$\sqrt{\frac{\hat{\sigma}_2{}^2 + b^2\hat{\sigma}_1{}^2 - 2\hat{\rho}b\ \hat{\sigma}_1\hat{\sigma}_2}{N}} > \sqrt{\frac{\hat{\sigma}_2{}^2 + \hat{\sigma}_1{}^2 - 2\hat{\rho}\hat{\sigma}_1\hat{\sigma}_2}{N}} \quad .$$

Obviously, if $b < 1$, these inequalities reverse. Thus, whether the norm-referenced test is conservative or liberal depends on the ratio of the norm group's spring standard deviation to its fall standard deviation. When $b > 1$, the denominator of the t test is too small and will give a larger Type I error rate than that reported.

Table 2-2 shows the values of b for the MAT subtests used in the PIP evaluation. For these subtests, b tends to be about 1.05, with a minimum value of about 0.8 and a maximum value of about 1.2. These values imply that, if the observed variances are nearly equal and if the observed correlation is about 0.5, the denominator for the t statistic obtained from the norm-referenced procedure may be anywhere from 8% too large to 11% too small. On the whole, however, this error may not be important, compared with the error introduced by uncritically taking the nonlongitudinal MAT standard score norms as valid estimates of fall-spring growth.

2.3.3 Criticisms of the Criterion of Educationally Significant Growth

We have mentioned that one-third standard deviation above normal growth is the criterion for educational significance. For the norm-referenced analysis, it is convenient to refer to this as "the criterion." However, in the PIP analysis it is not "one" criterion because, as shown in Table 2-2, the standard deviation of the norm group changes as a function of time and subtest.

Because the criterion of significance changes, the relationship between it and plausible measures of growth is not fixed, but may vary somewhat. For example, since average growth decreases with grade, if we compare the criterion for significant growth with the gain at the 50th percentile, we find that $(1/3)\sigma(t)$ is an increasing fraction of expected growth, unless $\sigma(t)$ decreases as a function of time. Table 2-2 shows that σ tends to increase with time, but the pattern is not consistent because of decreases in the middle grades. However, on balance, educationally significant growth is a larger fraction of 50th percentile growth at higher grades. It is not clear that we want our criterion for growth to be harder to obtain, relative to mean change, at higher grades.

To assess the relationship between $\sigma(t)$ and the expected growth on Total Reading for various percentiles, a polynomial regression equation

Table 2-2

RATIO OF SPRING TO FALL STANDARD DEVIATIONS FOR NATIONAL STANDARDIZATION GROUPS ON SELECTED MAT SUBTESTS USED IN THE PIP EVALUATION

Grade	Word Knowl- edge	Reading	Total Reading	Math Compu- tation	Math Con- cepts	Math Problem Solving	Total Math (Numbers)	$\bar{X}$	σ^2
Grade 2									
Fall SD	11.4	11.6	10.9				12.1		
Spring SD	10.2	11.7	10.9	10.0	11.8	12.1	11.1		
b	0.89	1.01	1.00				0.92	0.96	0.06
Grade 3									
Fall SD	11.4	11.7	11.6	10.2	11.8	12.7	11.4		
Spring SD	12.0	13.6	13.0	10.7	12.4	13.1	12.0		
b	1.05	1.16	1.12	1.05	1.05	1.03	1.05	1.07	0.05
Grade 4									
Fall SD	12.8	14.5	14.0	10.7	12.5	13.1	12.0		
Spring SD	12.8	14.5	14.3	12.1	12.0	13.3	12.1		
b	1.00	1.00	1.02	1.13	0.96	1.02	1.01	1.02	0.05
Grade 5									
Fall SD	12.7	14.0	13.5	9.9	11.3	12.5	10.4		
Spring SD	13.0	12.4	13.0	11.2	12.7	13.0	12.2		
b	1.02	0.89	0.96	1.13	1.12	1.04	1.17	1.05	0.1
Grade 6									
Fall SD	13.6	15.3	14.7	11.6	12.1	13.9	12.1		
Spring SD	13.7	12.6	13.5	12.2	14.3	13.6	12.7		
b	1.01	0.82	0.92	1.05	1.18	0.98	1.05	1.00	0.04
Grade 7									
Fall SD	13.5	15.1	14.7	11.3	11.9	13.7	12.2		
Spring SD	14.5	16.1	15.8	12.7	12.5	14.2	12.8		
b	1.07	1.07	1.07	1.12	1.05	1.04	1.05	1.07	0.01
Grade 8									
Fall SD	15.0	15.9	16.1	13.1	13.6	14.7	13.6		
Spring SD	15.5	17.1	16.9	14.4	14.8	15.0	14.4		
b	1.03	1.08	1.05	1.10	1.09	1.02	1.06	1.06	0.03
Grade 9									
Fall SD	15.7	16.4	16.8	14.7	15.2	14.7	14.2		

Note: b = spring divided by fall.
SD = standard deviation.

Source: Special Report No.8, Harcourt Brace Jovanovich, Inc., June 1971 was fit to the normative standard scores and percentiles. The basic idea is to simultaneously find two polynomials, $P_1(t)$ and $P_2(t)$ so that for a standard score y:

$$f(z,t) = y = P_1(t) + zP_2(t) \quad ,$$

where z is the normal deviate corresponding to the percentile for y. In this representation, P_1 describes changes at the mean as a function of time, and P_2 describes changes in the standard deviation.

After some preliminary runs, a second degree polynomial was selected for P_1, and a fifth degree for P_2. The resulting equations are:

$$\mu(t) = P_1(t) = \left[0.17 + \left(8.2 \times 10^{-3}\right) t - \left(2.8 \times 10^{-5}\right) t^2\right] \times 132$$

$$\sigma(t) = P_2(t) = \left[0.12 - \left(7.4 \times 10^{-3}\right) t + \left(3.7 \times 10^{-4}\right) t^2 - \left(7.3 \times 10^{-6}\right) t^3 + \left(6.3 \times 10^{-8}\right) t^4 - \left(2 \times 10^{-10}\right) t^5\right] \times 132 \quad ,$$

where t is time in months from beginning of kindergarten.

We have shown the coefficients to two places. The five place equation we fit by BMD 07R* has a coefficient of determination of 0.995 on 756 error degrees of freedom. Although this coefficient is large enough for our present purposes, the reader is cautioned that errors as large as 10% can be found fairly frequently, when predicted norm standard scores are compared with actual scores. Overall, however, predicted standard scores for the five place equation are quite close to the actual scores found in the norm tables. Presumably, this good fit reflects the normalizing transformation used to construct the standard scores.

Table 2-3 shows the rates of change of the mean and standard deviation for the fitted normative standard scores for Total Reading. Compared with changes in the mean, changes in the standard deviation are small. The fitted data indicate there is a slight tendency for the standard deviation to decrease at the higher grades; essentially, however, the criterion for growth--one-third the MAT norm standard deviation--is a constant criterion for all grades. Since average growth decreases as a function

---

*Biomedical Computer Programs (1973), pp. 387, ff, as modified by George Byrd of SRI.

Table 2-3

RATE OF CHANGE OF MEAN AND STANDARD DEVIATION
FOR FITTED STANDARD SCORES: TOTAL READING

Grade	Months from Kinder-garten	Monthly Rate of Change of μ (1)	Rate of Change of μ per Nine Months (2)	Monthly Rate of Change of σ (3)	Rate of Change of σ per Nine Months (4)	Ratio of (3) to (1) (5)
1	14	0.979	8.81	-0.090	-0.81	-0.091
	20	0.935	8.42	0.065	0.59	0.071
2	26	0.890	8.01	0.133	1.20	0.150
	32	0.846	7.61	0.140	1.26	0.170
3	38	0.802	7.22	0.111	1.00	0.14
	44	0.757	6.81	0.063	0.57	0.083
4	50	0.713	6.42	0.013	0.12	0.019
	56	0.668	6.01	-0.029	-0.26	-0.043
5	62	0.624	5.78	-0.056	-0.50	-0.087
	68	0.560	5.04	-0.065	-0.59	-0.117
6	74	0.535	4.82	-0.057	-0.51	-0.106
	80	.491	4.42	-0.039	-0.35	-0.079
7	86	0.447	4.02	-0.019	-0.17	-0.042
	92	0.402	3.62	-0.012	-0.11	-0.030
8	98	0.358	3.22	-0.035	-0.32	-0.099
	104	0.314	2.83	-0.110	-0.99	-0.350
9	110	0.269	2.42	-0.262	-2.36	-0.975

of grade, the criterion is an increasing function of average growth. As already mentioned, the implication is that in terms of average growth it is harder to achieve the criterion at higher grades.

The following question naturally arises: If the criterion is not a constant fraction of average growth, is it--for a given grade--a constant fraction of the growth at other percentiles? To answer this question, we refer to our model and evaluate:

$$\frac{(1/3)P_2(t)}{\frac{dP_1(t)}{dt} + \frac{zdP_2(t)}{dt}}$$

for selected values of t, where z is the normal deviate that corresponds to a percentile being investigated. Table 2-3 shows $dP_1(t)/dt$ and $dP_2(t)/dt$ for selected values of t in Columns (1) and (3), respectively.

Table 2-4 shows

$$\left\{9\left[\frac{dP_1(t)}{dt} + \frac{zdP_2(t)}{dt}\right]\right\}^{-1}$$

for the second and eighth grades. We are neglecting $(1/3)P_2(t)$, since it is only a scale factor. Table 2-4 shows that the ratio of the criterion to the growth of the fitted data is relatively constant across percentiles within grades, but not between grades. However, the relatively little variation within grades should probably not be neglected because at the second grade the criterion is a larger fraction of growth of the fitted scores at low percentiles than at high ones, while the reverse is true at the eighth grade. In column (3) of Table 2-4, we see that the main explanation is

Table 2-4

RECIPROCAL OF THE RATE OF CHANGE
OF FITTED STANDARD SCORES,
FOR SELECTED PERCENTILES
AND GRADES: TOTAL READING
(Per Nine Months)

Percentile	Spring, Second Grade (1)	Spring, Eighth Grade (2)	Ratio of (2) to (1) (3)
10	0.167	0.244	1.461
20	0.153	0.273	1.784
30	0.144	0.299	2.076
40	0.137	0.325	2.372
50	0.131	0.353	2.695
60	0.126	0.387	3.071
70	0.121	0.432	3.570
80	0.115	0.500	4.348
90	0.108	0.640	5.926

that, for the fitted data, Total Reading for children at the 90th percentile of the norm group in the second grade is growing almost 6 times faster than Total Reading of the eighth grade norm children at the same percentile, while Total Reading of second grade children at the 10th percentile is growing only about 1.5 times faster.

Thus, we find that the detailed answer to the question (of whether, for a given grade, the criterion is easier to achieve relative to some percentiles than to others) must be given on a grade-by-grade basis. We have found two grades at which the answer is affirmative. It is not clear that one would desire a criterion of educationally significant growth to be more difficult to obtain relative to some percentiles than to others, especially if the percentile that is more difficult depends on grade.

However, it is important to note that the nonconstancy of the criterion just reflects changes in the norms. Since the norms are not longitudinal, they are not strong evidence that the rate of growth for given percentiles changes as a function of grade, or that the ranking of rates for percentiles reverses for selected grades. It may be that the phenomena are just apparent and that a constant criterion is, in fact, reasonable. Nevertheless, as already remarked, the norm-referenced analyses assumes these phenomena are not apparent.

As discussed, there are several technical problems with the norm-referenced analysis, but the technical problems are not the reasons we reject the norm-referenced point of view. Our opinion is that the technical problems are just symptoms, as discussed in Section 2.4, which describes the analysis we prefer.

Further discussion of the properties of norm-referenced analysis can be found in Section 3.8 and in a forthcoming SRI technical publication authored by Kaskowtiz et al. Several of the results presented above were first obtained by him.

2.4 Justification for the Curriculum-Referenced Evaluation

Broadly speaking, the purpose of evaluating an educational innovation is to see whether the innovation meets selected objectives. Often the evaluator must choose the objectives that the innovation is to meet. He may choose from the objectives of federal policymakers or of the innovator, or may choose those implied by his own values.

We admit that our decision to reject the norm-referenced model is based on our concept of what is a reasonable evaluation of a field experiment like the PIP field trials. The decision is based on our values, not on technical points. For the convenience of the reader, the next paragraphs state our biases and their implications for the design of evaluations.

We propose as a principle that an evaluation must describe what occurred. An adequate evaluation does this in enough detail to assure reasonable men that there is a direct connection between the processes of interest and the outcomes of interest. Our principle demands only that the major terms in the evaluation be defined in enough detail so that their relationships can be seen. For the PIP field experiment, this means that "PIP," "MAT raw score," and related terms must be explained in language that allows competent educators to know what is meant. Obviously, there is nothing profound here; we are merely making our view explicit.

Relative to the definition of technical terms, we propose the positivist's "principle of abstraction." This principle is a version of Occam's Razor: Entities should not be increased without reason. The principle of abstraction states that whenever one desires to define an entity that is said to be common to a collection of entities, it is sufficient to refer to the collection. (This eliminates the abstract property; the principle of abstraction is actually a principle that does away with abstraction.)

No merely formal analysis of evaluation data will satisfy our proposed evaluation principles because a merely formal analysis will leave the main treatment and outcome terms operationally defined (e.g., "achievement" would be defined as whatever the MAT measures). In this evaluation, the norm-referenced analysis, as its name suggests, is completely dependent on the operational interpretation of "achievement" as whatever the MAT norms measure. This interpretation is related to the concept of "true score," which is operationally defined as whatever the items that were answered correctly have in common. "True scores" defined in this way is just the sort of technical term that the positivist's version of Occam's Razor eliminates. We feel that the technical troubles associated with the norm-referenced analysis are symptoms of the confusion that results when we allow ourselves to focus on entities, like true scores, that would be eliminated by the principle of abstraction. Without assuming the existence of true scores, the Thurstone technique, which make the cross-sectional norms appear longitudinal, loses its appeal.

The principle of abstraction is quite radical, unless taken with a grain of salt. Applied to technical and to nontechnical terms, it would eliminate so many entities that ordinary conversation would be impossible. The thing called "wall paper," for instance, would be eliminated; we would have to refer to its more elemental properties, like its color and extent.

Obviously, we do not wish to eliminate the conveniences of ordinary linguistic conventions. The principle of abstraction should be applied in making the definitions of key technical terms, but not applied to the point where we lose the ability to communicate plainly.

When these ideas are used in defining this evaluation's principal dependent measure, the MAT raw score, we are compelled to view the MAT not as an entity, but as a collection of items. When we define "PIP," we are compelled to view a PIP not as an entity, but as a set of instructions for implementing a specific project. To satisfy our first principle, we are obligated to display those instructions that are supposed to make children answer the MAT items correctly, and we must supply reasons for believing that responses to the MAT items are connected to the PIP instructions.

The connection we assert is that, all other things being present, items will be learned from a competent teacher, given an appropriate curriculum and a reasonable length of exposure to it. Thus, our evaluation of the PIPs <u>as they impact achievement test scores</u>, is aimed at determining the connections between the PIP-specified curriculum that was used, instructional procedures that were used, and the MAT items.

If we find that MAT items (that were covered by the PIP-specified curriculum materials that were used) are not learned, we would have no evidence of PIP success.

If we find that MAT items (that were covered by the PIP-specified curriculum materials that were used) are learned, we would have evidence consistent with success, but we would have no proof of success. This is because all PIP projects, except R-3, are "pull-out" programs in which the children spend only a few hours a day. Therefore, the regular school curriculum may be responsible for any observed gains.

The curriculum-referenced analysis, therefore, has the following distinct activities:

- An in-depth examination of each PIP's curriculum and its instructional techniques.
- A program of data collection to verify that the curriculum and instructional techniques were used in the field.
- An analysis of the curriculum that was PIP-specified and used, with a determination of which items of the MAT are relevant to the curriculum observed in the field.
- Data analyses that quantitatively assess the association between PIP implementation variables and relevant items.

2.5 Report Organization

In spite of the earnest preachings of those who subscribe to our evaluation principles, some readers will assert that true scores are a useful construct. For them, and to complete the black-box analysis described in Section 2.1, we give in Section 3 the results of the norm-referenced analysis and the results of a modified norm-referenced analysis. The modified version attempts to overcome the technical problems that were discussed earlier, as well as several additional technical problems.

Section 4 introduces the curriculum-referenced analysis with a discussion of the activities that were directed toward defining the PIP instructional programs and their curricula. It also describes the results of the field activities that were geared to discovering wha. instructional components were implemented.

Section 5 describes our analysis of the relationship of the MAT items to the curriculum that was PIP-specified and used. It also investigates the issue of whether the tests given to the originating sites were more closely related to the PIP curriculum than the MAT is. We also discuss the possibility of analyzing children's scores with the discussion limited to only those items that were shown to be covered in the PIP program.

Section 6 consists of formal analyses relating PIP-relevant MAT items to implementation variables.

3 RESULTS OF THE NORM-REFERENCED ANALYSIS

3.1 Introduction

In Section 3, we present the results of the norm-referenced analysis of the MAT data. The norm-referenced analysis is regarded by some as the principal measure of PIP effectiveness. That is, if the PIP design is successful, the PIP projects will show educationally significant gains for their students, just as the originating sites did. We have examined the generic true-score justification for the norm-referenced analysis and have found that it had no empirical content; even though several achievement tests were used in selecting the originating sites, no actual sampling of tests or students was done. Since this evaluation is based on a single test, the MAT, the following analyses are not operationally equivalent to those conducted on the originating sites. For that reason our analyses cannot, strictly speaking, show the results that the originating sites did. Because of this nonequivalence, as well as because of our distrust of the longitudinal validity of the MAT standard-score norms, we do not feel that the results shown below adequately test PIP effectiveness.

3.2 Test Selection

Although strict adherence to the concepts of generic true-score theory implies that one need not be overly concerned about the details of test selection, no one recommends that the generic theory be taken that seriously.

Because the norm-referenced procedures depend on having credible norms, we took care to select a test that had empirical norms for both fall and spring. The MAT has this feature. We also took care to test in October and April, the months in which the MAT normative testing was done. Only later did we realize that only fall data were used in the equivalencing of test batteries.* The use of only part of the data

---

*The total sample of students at each grade level taking Form G of the MAT during the fall standardization was used as the scaling population (MAT Guidelines No. 1, 1972).

probably accounts for the erraticism shown in Figure 2-1 and possibly accounts for the slight reduction in norm standard-score variance at the middle grades.

We also required that the test permit out-of-level testing, since many of the PIP children were thought to be at least one grade level below their normal grade. Through the device of standard scores, the MAT also had this feature. However, use of standard scores is not equivalent to using the "raw" norms directly. We have already observed that the standard scores appear quite smooth statistically (as opposed to graphically). It was observed that only 9 parameters account for essentially all of the variance in the 16 norm tables for MAT Total Reading. It is difficult to imagine how we could predict "raw" norms that well with the same number of parameters.

Since the test selection was made before the PIPs were operational, the curriculum that would be employed in the field could not be determined. However, we could see that, roughly speaking, the MAT items covered a good range of topics in mathematics and that the reading items were reasonable.

Overall, we feel that the MAT is one of the best off-the-shelf achievement test batteries. It certainly satisfies the requirements of generic true-score theory, and, for those who are not put off by the consequences of assuming that the MAT norms are valid longitudinally and known without error, the MAT provides a very good basis for a norm-referenced analysis.

So that those who are unfamiliar with the MAT battery can understand the types of items designated by the subtest labels (such as "Mathematics Concepts"), we describe below the MAT subtests used in the PIP evaluation.

- MAT Primer--The MAT Primer was given in the fall to all first grade children in the study (Catch-Up, Conquest, PTR) and in the spring to the first grade children in Canton (PTR). The Primer is composed of three subtests: Listening for Sounds, Reading, and Numbers. The Numbers subtest was not applicable to PTR or Conquest, which are reading programs, and was therefore not administered to children in those projects.

  The Listening for Sounds subtest contains 39 items. Twenty-one items require the child to match a sound spoken by the tester with a picture of an object whose name begins or ends with that sound. Eight items require the child to match a spoken word with a written word.

The Reading subtest consists of 33 items. In 11 items the child must select a letter that the tester has said aloud from a group of four letters. Seventeen items require the child to select the one word out of four that "best tells" about a picture. The remaining five items require the child to select one of three sentences that best describes a picture.

The Numbers subtest is composed of 34 items. Twenty items are read to the child. These items test the child's knowledge of shapes, sizes, 1:1 correspondence, numerical recognition, money, measurement, time, place value, and number series. Fourteen items require the child to do some simple one-digit addition and subtraction problems and write the answer in the test booklet.

- MAT Primary I--The MAT Primary I was given to all second grade students during the fall (Catch-Up and Conquest). In the spring it was given to all the first graders in the study (Catch-Up, Conquest, and PTR), except those in the Canton PTR project. This test consists of four subtests: Word Knowledge, Word Analysis, Reading, and Mathematics. The Word Analysis subtest was given only in the spring to the first graders. The Mathematics subtest was given only at the Catch-Up sites.

The Word Knowledge subtest consists of 35 items; for each, the child is required to select from among four words the one word that best describes a picture.

The Word Analysis subtest is composed of 40 items; for each, the child must match a spoken word with a written word.

The Reading subtest has two parts, consisting of 13 and 29 items, respectively. In the first part of the subtest the student selects one of three sentences that best describes a picture. In the second part he must answer eight riddles and then read five simple paragraphs and answer questions about each one.

The Mathematics subtest is also divided into two parts. The first 35 items examine the student's understanding of counting, money, measurement, place value, and story problems. The second part, which consists of 27 items, tests the student's ability to add and subtract one- and two-digit numbers and to solve some simple equations, such as: $4 + \_ = 7$.

- MAT Primary II--The MAT Primary II was given to all third graders in the fall (Catch-Up, Conquest, and IRIT) and to all second graders in the spring (Catch-Up and Conquest). Primary II consists of seven subtests, five of which were administered to the PIP students: Word Knowledge, Reading, Mathematics Computation, Mathematics Concepts, and Mathematics Problem Solving. The mathematics subtests were given only to students in Catch-Up, since Conquest and IRIT are reading programs.

  The Word Knowledge subtest consists of 40 items. The first 17 require the student to select from four words the one that best describes a picture. The remaining items require the student to identify synonyms and antonyms.

  The Reading subtest is also divided into two parts. The first 13 items require the student to choose one of three sentences that best describes a picture. The remaining 31 items are questions about six simple paragraphs that the student must read.

  Mathematics Computation, a subtest of 33 items, requires the student to add one- and two-digit numbers, with two and three addends, multiply one-digit numbers, and solve simple equations, such as: $28 - \_\_ = 19$.

  The Mathematics Concepts subtest, as in Primary I, tests the student's knowledge of geometry, measurement, concepts of fractions, place value, number series, inequality, and properties of number systems.

  The Mathematics Problem Solving subtest consists of 35 items, about one-half of which are dictated by the tester. All items are simple story problems, with the exception of two that instead require the child to pick the correct number sentence from a group of four.

- MAT Elementary--The MAT Elementary was administered to all fourth, fifth, and sixth graders during the fall (Catch-Up, Conquest, HIT, and IRIT). In the spring, the Elementary was given to all third and fourth graders (Catch-Up, Conquest, and IRIT). The Elementary consists of the same seven subtests as the Primary II, five of which were administered to the students: Word Knowledge, Reading, Mathematics Computation, Mathematics Concepts, and Mathematics Problem Solving. The three math subtests were administered to Catch-Up and HIT in the fall and only to Catch-Up in the spring.

The Word Knowledge subtest contains 50 items that require the student to identify synonyms and antonyms.

The Reading subtest is made up of 45 items that require the student to read some stories and to identify the main idea, draw inferences, and determine word meanings from the context.

Mathematics Computation is a subtest of 40 items that require the student to perform addition, subtraction, multiplication, and division. Seven items require the manipulation of decimals and fractions. Four items are simple mathematical sentences like $\_\_\_ \div 9 = 9$.

The Mathematics Concepts subtest contains 40 items that attempt to assess the student's understanding of basic mathematical principles and geometry.

The Mathematics Problem Solving subtest consists of 35 mathematics word problems. Three items require knowledge of how to read a chart.

- MAT Intermediate--The MAT Intermediate was given to all seventh grade students in the fall (Catch-Up and HIT). In the spring it was given to all fifth and sixth grade students (Catch-Up, Conquest, and HIT). The format of the Intermediate parallels that of the Elementary, and the same five subtests were administered. Again, students in Conquest did not take the mathematics subtests because Conquest is a reading program.

  The Word Knowledge and Reading subtests have the same number of items and the same format as the Elementary, although the questions tend to be somewhat more difficult.

  The three mathematics subtests in the Intermediate have more items that focus on fractions, percents, decimals, and rounding than does the Elementary.

- MAT Advanced--The MAT Advanced was given to all eighth and ninth grade students in the fall (Catch-Up, HIT, and R-3). In the spring it was given to all seventh, eighth, and ninth grade students (Catch-Up, HIT, and R-3). The same five subtests were given. The format of the Advanced parallels that of the Intermediate, and the same numbers of items are included in each of the subtests, but the difficulty factor has been increased.

3.3 Test Scheduling and Administration

The major features of the test scheduling were determined by our desire to test syncronously with the MAT norms and by the schedules of the schools that housed the PIP projects.

The desire for staying within the MAT norm testing dates (mid-October and mid-April) meant that only the middle cycle of IRIT projects could be tested. Since there were no MAT norms above the ninth grade (and the ninth grade norms are extrapolated), we decided not to test the tenth grade PIP participants in the Olean HIT project.

Fall testing was completed between 6 October and 24 October 1975 by "test teams" composed of a test administrator and a test monitor. Every available PIP participant was tested. Spring testing was completed between 5 April and 7 May 1976. PIP participants who were absent during the fall testing period, or whose tests were subsequently invalidated, were not tested in the spring. The rather long testing period in the spring was caused by the necessity to accommodate local testing plans and Easter vacation.

We attempted to complete testing within five working days at each PIP project, both in fall and in spring. Table 3-1 shows the fall and spring testing dates for each project, and the number of test teams.

In those projects where testing could be completed within five working days by one test team, the SRI site visitor served as test administrator and the local site assistant served as test monitor. Where testing could not be completed within five working days by one test team, local personnel, in addition to the site assistant, were hired and trained by the site visitor to serve as test administrators and test monitors. One exception was Benton Harbor, where testing could not be completed within the designated time by one test team and where local conditions did not provide for proper use of more than one test team. An SRI floating site visitor assisted the site visitor assigned to Benton Harbor.

The test teams were trained for one or two days by the SRI site visitor in accordance with the Manual of Procedures for Project Information Packages Testing. They were then supervised each day by the site visitor throughout the duration of the testing.

The Manual of Procedures for Project Information Packages Testing was developed to be the reference manual for the achievement testing phase of the PIP evaluation. This manual explained the duties of the

Table 3-1

TEST DATES AND NUMBER OF TEST TEAMS

Project	Fall 1975 Test Dates	Fall 1975 Number of Test Teams	Spring 1976 Test Dates	Spring 1976 Number of Test Teams
Benton Harbor	10/6-10/9	1.5*	5/3-5/7	1.5*
Bloomington (Catch-Up)	10/9-10/15	3	4/15-4/21	3
Bloomington (IRIT)	10/15-10/17	1	4/21-4/22	1
Brookport	10/9, 10/10, 10/14	1	4/6-4/9	1
Canton	10/6-10/9	4	4/12-4/15	4
Charlotte	10/6-10/9	4	4/12-4/15	3
Cleveland	10/14-10/17	6	4/7-4/8 4/12-4/14	5
Dallas	10/13-10/16	3	4/5-4/8	3
Galax	10/20-10/22	1	4/12-4/14	1
Gloversville	10/20-10/23	2*	4/26-4/29	2*
Lake Village	10/13-10/16	1	4/19-4/22	1
Lexington	10/6-10/10	3	4/12-4/15	3
Lorain	10/13-10/16	4	4/20-4/23	3
Oklahoma City	10/6	1	4/12	1
Olean	10/20-10/24	3	4/5-4/8	3
Providence Forge	10/16-10/17	1	5/3-5/4	1
Schenectady (IRIT)	10/21-10/22	1	4/27	1
Schenectady (R-3)	10/17-10/19	3	4/29	9
Wayne City	10/6-10/8	1	4/6-4/8	1

*Indicates more than one SRI site visitor.

site visitor, test administrator, test monitor, and site assistant. It also described the procedures for packing and shipping completed tests to SRI. The spring version of the manual is shown in Appendix A.

As shown in Table 3-1, the number of test teams for three projects (Cleveland, Charlotte, and Lorain) decreased by one in the spring. We had decided to test in the spring only those children with valid fall tests, which reduced the number of test sessions required. In the Schenectady R-3 project, the number of test teams was increased from three to nine in the spring, and the number of days of testing was reduced from three to one. This was done to accommodate local test schedules and to ensure cooperation from the host school district.

The test battery for both fall and spring consisted of the MAT (Form F) and one of two affective tests--either the Faces Attitude Inventory or the Intellectual Achievement Responsibility Scale (IAR). In addition to the above test battery, a PIP and site-specific Student Attitude Questionnaire (SAQ) was administered in the spring. A discussion of the student attitude measures appears in Appendix A to Volume One. Since five of the six PIPs were designed to supplement reading and/or math curriculum, only the reading and math tests of the MAT were given. Although the sixth PIP (R-3) was a replacement program (that is, replaced the entire curriculum), only the reading and math tests were administered because the criteria for effectiveness at the originating site dealt only with reading and math. Table 3-2 shows the grade levels tested at each project and the test administered.

Table 3-2 also shows the MAT levels assigned to each grade level in the fall and spring. With the exception of grades 5, 6, and 7, all fall tests administered were at levels recommended by the test publisher. Originally the plan was to use the same test at each grade level for the pre- and post-test. However, when the fall 1975 test scores showed no serious "bottoming out" in terms of raw scores, a decision was made to move to the recommended test levels for all grades. This decision was made because of our desire to stay close to the empirical norms and to prevent ceiling effects from obscuring PIP effects. An exception was made for the Canton PTR project. There is no mandatory kindergarten in the State of Mississippi, hence the PTR students were in their first year of school. Examination of each student's placement in the PTR curriculum in early spring 1976 indicated that none of the PTR students would hit the ceiling of the Primer, while moving up to the Primary I level would have been unfair to many students, who simply would not comprehend the Primary I questions.

Table 3-2

PIP TEST PLAN

a. Fall 1975

Project	Metropolitan Achievement Tests																		Affective Tests	
	Primer		Primary I		Primary II		Elementary						Intermediate		Advanced				FACES	IAR
	Grade 1		Grade 2		Grade 3		Grade 4		Grade 5		Grade 6		Grade 7		Grade 8		Grade 9		Grades	Grades
	Read	Math	Read	Math	Read	Math	Read	Math	Read	Math	Read	Math	Read	Math	Read	Math	Read	Math	1-2	3-9
Catch-Up																				
Bloomington	X	X	X	X	X	X	X	X	X	X									X	X
Brookport	X	X	X	X	X	X	X	X	X	X	X	X							X	X
Galax	X	X	X	X	X	X	X	X	X	X	X	X							X	X
Providence Forge							X	X	X	X	X	X								X
Wayne City	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X			X	X
Conquest																				
Benton Harbor			X		X		X		X		X								X	X
Cleveland	X		X		X		X		X		X								X	X
Gloversville			X		X		X		X		X								X	X
HIT																				
Lexington											X	X	X	X	X	X	X	X		X
Olean													X	X	X	X	X	X		X
IRIT																				
Bloomington					X		X													X
Oklahoma City					X															X
Schenectady					X		X													X
PTR																				
Canton	X																		X	
Dallas	X																		X	
R-3																				
Charlotte															X	X				X
Lake Village															X	X				X
Lorain															X	X				X
Schenectady															X	X				X

53

Table 3-2 (Concluded)

b. Spring 1976

Project	Metropolitan Achievement Tests																			Affective Tests		Student Attitude Questionnaires	
	Primer	Primary I		Primary II		Elementary				Intermediate				Advanced						FACES	IAR	FSAQ	CSAQ
	Grade 1	Grade 1		Grade 2		Grade 3		Grade 4		Grade 5		Grade 6		Grade 7		Grade 8		Grade 9		Grades	Grades	Grades	Grades
	Read	Read	Math	Read	Math	Read	Math	Read	Math	Read	Math	Read	Math	Read	Math	Read	Math	Read	Math	1-2	3-9	1-2	3-9
Catch-Up																							
Bloomington		X	X	X	X	X	X	X	X	X	X									X	X	X	X
Brookport		X	X	X	X	X	X	X	X	X	X	X	X							X	X	X	X
Galax		X	X	X	X	X	X	X	X	X	X	X	X							X	X	X	X
Providence Forge								X	X	X	X	X	X								X		X
Wayne City		X	X	X	X	X	X	X	X	X	X	X	X	X	X	X				X	X	X	X
Conquest																							
Benton Harbor				X		X		X		X		X								X	X	X	X
Cleveland		X		X		X		X		X		X								X	X	X	X
Gloversville				X		X		X		X		X								X	X	X	X
HIT																							
Lexington												X	X	X	X	X	X	X	X		X		X
Olean														X	X	X	X	X	X		X		X
IRIT																							
Bloomington						X		X													X		X
Oklahoma City						X															X		X
Schenectady						X		X													X		X
PTR																							
Canton	X																			X		X	
Dallas		X																		X		X	
R-3																							
Charlotte																X	X				X		X
Lake Village																X	X				X		X
Lorain																X	X				X		X
Schenectady																X	X				X		X

77

78

The MAT reading and math tests were administered in conformity with directions outlined in the Teacher's Directions booklet for each level of the test. These directions were incorporated into SRI's Manual of Procedures for Project Information Packages Testing. The manual also provided specific schedules for administering the reading or math tests, or a combination of both, at each level. Overall, the schedules were truncated from the schedules recommended by the publisher, since we administered only reading and math tests. We wished to administer the tests in as short a time as possible to reduce the time lost from instruction and to prevent student fatigue. Tests were administered over a period of one to three days, with the reading test given the first day. Tests were given in no more than three sittings, with an hour between sittings. Students receiving the Primer were tested in groups of 15 or fewer. Students receiving the remaining levels of the MAT were tested in regular class sizes, not to exceed 30 students.

Table 3-3 provides a comparison between the recommendations provided by the publisher and the schedules devised by SRI for completing the testing at each level. The table indicates some modifications in the order in which subtests were administered. In a few cases, the Reading subtest was administered before the Word Knowledge subtest to provide more efficient scheduling and to allow the most difficult subtest to be given to the lower grades when the students were the most alert. The table also shows that the Word Analysis subtest of the Primary I was administered in the spring to provide a comparison subtest for Listening to Sounds on the Primer for all first grades except Canton.

The practice items of each test were administered before actual testing, which added five to ten minutes to the first sitting for the Primer, Primary I, Primary II, and Elementary levels. Although the schedule provided in Table 3-3 was the same for both fall and spring, it does not show the Student Attitude Questionnaire, which was administered in the spring. The SAQ was administered in conjunction with the affective test in each case and added five to seven minutes to administration time.

3.4 Quality Control Procedures

To ensure a high degree of confidence in the field-test data, SRI incorporated quality control procedures during all phases of fall 1975 and spring 1976 testing. Quality control began with the shipment of test materials and ended with the test data on computer files.

Table 3-3

TEST SCHEDULES FOR FALL AND SPRING

	Publishers' Recommended Schedules			SRI Schedules: Total Test			SRI Schedules: Reading Subtest Only			SRI Schedules: Math Subtest Only		
Test	Sitting	Part	Time*	Sitting	Part	Time*	Sitting	Part	Time*	Sitting	Part	Time*
Primer	1	Practice page	10	1	Practice page	10	1	Practice page	10			
	2	Part 1: Listening for Sounds	20		Listening for Sounds	20		Listening for Sounds	20			
	3	Part 1: (Cont'd) page 4	5	2	Listening for Sounds (cont'd)		2	Listening for Sounds				
	4	Part 2: Reading, page 5	5		page 4	5		(cont'd) page 4	5			
	5	Part 2: (Cont'd) pages 6 & 7	[illegible]		Reading	20		Reading	20			
	6	Part 3: Numbers, pages 8 & 9			Numbers	25	3	Affective	15			
	7	Part 3: (Cont'd) page 10			Affective	15						
	8	Part 3: (Cont'd) page 11										
Primary I	1	Test 1: Word Knowledge			Test 3: Reading	30	1	Test 3: Reading	30			
		Test 2: Word Analysis		2	Test 1: Word Knowledge	15	2	Test 1: Word Knowledge	15			
	2	Test 3: Reading	[illegible]		Test 2: Word Analysis	15		Test 2: Word analysis	15			
	3	Test 4: Math Concepts	15		(Spring only)			(Spring only)				
		Math Computation	15									
					Affective	15		Affective	15			
				3	Test 4: Math Concepts	15						
					Math Computation	15						
Primary II	1	Test 1: Word Knowledge	18	1	Test 1: Word Knowledge	18	1	Test 3: Reading	30			
		Test 2: Word Analysis	15		Test 3: Reading	30	2	Test 1: Word Knowledge	18			
	2	Test 3: Reading	30	2	Test 5: Math Computation	18		Affective	15			
	3	Test 4: Spelling	10		Test 6: Math Concepts	20						
		Test 5: Math Computation	18	3	Test 7: Math Problem Solving	25						
	4	Test 6: Math Concepts	20		Affective	15						
	5	Test 7: Math Problem Solving	25									
Elementary	1	Test 1: Word Knowledge	15	1	Test 1: Word Knowledge	15	1	Test 2: Reading	25†	1	Test 5: Math Computation	35
	2	Test 2: Reading	25		Test 2: Reading	25	2	Test 1: Word Knowledge	15†		Test 6: Math Concepts	30
	3	Test 3: Language	30	2	Test 5: Math Computation	35		Affective	15†	2	Test 7: Math Problem	
	4	Test 4: Spelling	20		Test 6: Math Concepts	30					Solving	30
	5	Test 5: Math Computation	35	3	Test 7: Math Problem Solving	30	1	Test 1: Word Knowledge	15‡		Affective	15
	6	Test 6: Math Concepts	25		Affective	15		Test 2: Reading	25‡			
	7	Test 7: Math Problem Solving	30					Affective	15‡			
Intermediate	1	Test 1: Word Knowledge	15	1	Test 1: Word Knowledge	15	1	Test 1: Word Knowledge	15	1	Test 5: Math Computation	35
		Test 2: Reading	25		Test 2: Reading	25		Test 2: Reading	25		Test 6: Math Concepts	25
	2	Test 3: Language	50	2	Test 5: Math Computation	35		Affective	10	2	Test 7: Math Problem	
	3	Test 4: Spelling	15		Test 6: Math Concepts	25					Solving	25
		Test 5: Math Computation	35	3	Test 7: Math Problem Solving	25					Affective	10
	4	Test 6: Math Concepts	25		Affective	10						
		Test 7: Math Problem Solving	25									
Advanced	1	Test 1: Word Knowledge	15	1	Test 1: Word Knowledge	15	1	Test 1: Word Knowledge	15	1	Test 5: Math Computation	35
		Test 2: Reading	25		Test 2: Reading	25		Test 2: Reading	25		Test 6: Math Concepts	25
	2	Test 3: Language	45	2	Test 5: Math Computation	35		Affective	10	2	Test 7: Math Problem	
	3	Test 4: Spelling	25		Test 6: Math Concepts	25					Solving	25
		Test 5: Math Computation	35	3	Test 7: Math Problem Solving	25					Affective	10
	4	Test 6: Math Concepts	25		Affective	10						
		Test 7: Math Problem Solving	25									
	5	Test 8: Science	35									
	6	Test 9: Social Science	45									

*Excludes practice items except Primer.
†Grade 4 (fall), grades 3 and 4 (spring).
‡Grades 5 and 6 (fall).

80 81

For a description of quality control procedures in the field, see the Manual of Procedures for Project Information Packages Testing in Appendix A.

Formal procedures for handling materials at SRI assured quality control of coding, keypunching, and test scoring.

After the tests were returned to SRI, each test booklet within a test carton was examined, item by item, and coded for keypunching by trained personnel following specific instructions. This examination maintained quality control by ensuring that responses were clearly indicated and unambiguous for keypunching and that the contents of one test carton were not integrated with the contents of other test cartons. Care was taken that coding errors could be traced back to the individual responsible.

After the review and coding of each test item, each item response was made machine-readable through direct keypunching from the test booklets to disk. All keypunched documents were 100% key-verified and then transferred onto magnetic tapes. Each keypuncher identified the job(s) he completed and returned the test booklets and keypunched information to PIP project personnel. Keypunching was completed at SRI in the fall and subcontracted to an independent firm in the spring.

All test booklets were scored by a program written by project staff.* Our programs provided for built-in audit totals for the number of records processed. Edit checks were performed by computer to identify meaningless codes and, where appropriate, to test the data for logical inconsistencies. The edit checks were also performed to verify that all entries corresponded to the coding specifications and keypunch instructions. Computer tests for logical consistency dealt primarily with assuring that the correct sequence of subtests was being scored.

To ensure that field-test data were properly processed into machine-readable form, project staff manually examined at least a 10% sample of each type of test battery given, at least a 10% sample from each site, and at least a 10% sample of each test carton. This entailed comparing, item by item, each student's test booklets with the same information on the raw test data file. Any errors detected were flagged, and a notation was made next to each error identifying it as keypunch or coding error. Higher than average coding error rates indicated that all test booklets handled by the reviewer in question should be reexamined. This resulted in one item in the Elementary Reading subtest being completely rescored.

---

*Bert Laurence of SRI designed our procedures for producing scored records.

Error rates were calculated in terms of coding errors, keypunching errors, and overall number of errors. Since a data field on the keypunch cards was reserved for each response and since the data field sizes varied depending on the test battery, the total number of data fields manually checked was calculated. The total number of errors detected was then divided by the total number of fields checked to arrive at the error rate. Finally, all errors detected were corrected on the raw test data files. Table 3-4 shows the error rates encountered during quality control procedures for fall and spring testing.

Table 3-4

SUMMARY OF ERRORS FOUND DURING QUALITY CONTROL CHECK OF KEYPUNCHING AND CODING OF MAT SUBTESTS, FACES TESTS, AND IAR TESTS

	Fall	Spring
Sets of test booklets	3,491	3,491
Sets of test booklets sampled	359	349
Subtests checked	822	1,890
Subtest fields checked	26,752	83,686
Keypunch errors found	0	145
Error rate	0	0.0017
Coding errors found	9	9
Error rate	0.003	0.0001
Total error rate	0.003	0.0018

3.5 Invalidation of Tests

As described in Appendix A, the test administrator was allowed to invalidate subtests of PIP participants on site. Subtests were invalidated under one or more of the following conditions:

- Student refused to respond throughout most of the subtest
- Student borrowed answers consistently
- Student marked multiple answers consistently
- Student became ill during the subtest
- Student was absent
- Student worked the wrong subtest
- Student was in special education
- Student had a severe physical/mental handicap
- Other reasons specified.

The last condition provided for any unforeseen situation that, in the judgment of the test administrator, was serious enough to constitute invalidation.

Table 3-5 shows test invalidations, by type and number, for Total Reading and Total Math for each project in the fall. Table 3-6 provides the same information for the spring. Table 3-5 shows that, for both Total Reading and Total Math, most invalidations were due to absenteeism, followed by reasons categorized as "other." One project, Canton, had a significant number of invalidations in Total Reading because students provided multiple answers. We attribute this to their unfamiliarity with taking tests, since it was their first year of school.

Table 3-6 shows that, for both Total Reading and Total Math, a leading cause of spring invalidations was, again, absenteeism. However, most losses were due to the withdrawal of students from the PIP program. Children were classified as "withdrawn from the program" for several reasons. Some were simply noted as no longer in the program and were lost to us. In the IRIT projects, we listed the children not assigned to the middle cycle as withdrawn from the middle cycle; such children may have been assigned to other cycles. In Gloversville Conquest, we found that some children who were tested had not been assigned to the project full time; we also coded these children as withdrawn.

Because of various site conditions, we were unable to test a constant fraction of the children originally rostered in the projects. In IRIT, as already mentioned, the children in the first and third instructional cycles were not tested. In Gloversville, we were forced to invalidate 57 tests because of poor project implementation. Finally, poor test administration in Cleveland forced us to invalidate 37 tests in the fall.

Table 3-5

NUMBER OF INVALID TESTS, BY REASON FOR INVALIDATION — FALL

a. Total Reading

Project	Valid Tests		Reason for Invalidation: Refused Response		Borrowed Answers		Multiple Answers		Student Ill During Test	
	Number of Students	Percent of Rostered Students	Number of Students	Percent of Rostered Students	Number of Students	Percent of Rostered Students	Number of Students	Percent of Rostered Students	Number of Students	Percent of Rostered Students
Catch-Up										
Bloomington	171	89.06%								
Brookport	59	96.72								
Galax	59	92.19								
Providence Forge	61	95.31			1	1.56%				
Wayne City	41	97.62								
Total Catch-Up	391	92.43%			1	0.24%				
Conquest										
Benton Harbor	132	94.29%								
Cleveland	382	84.14	2	0.44%			2	0.44%		
Gloversville	213	96.82								
Total Conquest	727	89.31%	2	0.25%			2	0.25%		
HIT										
Lexington	238	97.94%								
Olean	217	94.35	1	0.43%						
Total HIT	455	96.20%	1	0.21%						
IRIT										
Bloomington	73	96.05%								
Oklahoma City	45	95.74								
Schenectady	[illegible]	93.62								
Total IRIT	162	95.29%								
PTR										
Canton	140	74.07%	20	10.58%	1	0.53%	15	7.94%		
Dallas	161	87.5								
Total PTR	301	80.70%	20	5.36%	1	0.27%	15	4.02%		
R-3										
Charlotte	[illegible]	92.86%	2	0.68%						
Lake Village	134	98.53								
Lorain	328	89.86								
Schenectady	194	87.39								
Total R-3	927	91.35%	2	0.20%						
Total, all projects	2,965	90.67%	25	0.76%	2	0.06%	17	0.52%	0	0

Table 3-5 (Continued)

Total Reading (Concluded)

Project	Reason for Invalidation: Absent — Number of Students	Absent — Percent of Rostered Students	Wrong Subtest — Number of Students	Wrong Subtest — Percent of Rostered Students	Special Education — Number of Students	Special Education — Percent of Rostered Students	Student Has Handicap — Number of Students	Student Has Handicap — Percent of Rostered Students	Other — Number of Students	Other — Percent of Rostered Students	Total Students Rostered
Catch-Up											
Bloomington	9	4.69%							12	6.25%	192
Brookport	2	3.28									61
Galax	5	7.81									64
Providence Forge	2	3.13									64
Wayne City	1	2.38									42
Total Catch-Up	19	4.49%							12	2.84%	423
Conquest											
Benton Harbor	8	5.71%									140
Cleveland	30	6.61	1	0.22%					37	8.15%	454
Gloversville	7	3.81									220
Total Conquest	45	5.53%	1	0.12%					37	4.55%	814
HIT											
Lexington	5	2.06%									243
Olean	9	3.91							3	1.30%	230
Total HIT	14	2.96%							3	0.63%	473
IRIT											
Bloomington	3	3.95%									76
Oklahoma City	2	4.26									47
Schenectady	3	6.38									47
Total IRIT	8	4.71%									170
PTR											
Canton	[illegible]	1.59%							10	5.29%	189
Dallas	22	11.69							1	0.54	184
Total PTR	25	6.70%							11	2.95%	373
R-3											
Charlotte	[illegible]	5.78%							2	0.68%	294
Lake Village	[illegible]	1.47									136
Lorain	27	7.40			6	1.64%			4	1.10	365
Schenectady	18	8.11			1	0.45			9	4.05	222
Total R-3	64	6.29%			7	0.69%			15	1.47%	1,017
Total, all projects	[illegible]	5.35%	1	0.03%	7	0.21%	0	0	78	2.39%	3,270

Table 3-5 (Concluded)

b. Total Math

Project	Valid Tests: Number of Students	Valid Tests: Percent of Rostered Students	Refused Response: Number of Students	Refused Response: Percent of Rostered Students	Borrowed Answers: Number of Students	Borrowed Answers: Percent of Rostered Students	Multiple Answers: Number of Students	Multiple Answers: Percent of Rostered Students	Student Ill During Test: Number of Students	Student Ill During Test: Percent of Rostered Students
Catch-Up										
Bloomington	178	92.71%								
Brookport	56	91.80								
Galax	59	92.19								
Providence Forge	58	90.63								
Wayne City	41	97.62								
Total Catch-Up	392	92.67%								
HIT										
Lexington	104	85.25%	1						1	0.82%
Olean	186	93.0	[illegible]	[illegible]					1	0.50
Total HIT	290	90.6[illegible]%	2	[illegible]					2	0.62%
R-3										
Charlotte	272	92.52%								
Lake Village	131	96.32								
Lorain	331	90.68								
Schenectady	190	85.59	[illegible]	0.45%						
Total R-3	924	91.86%	1	0.10%						
Total, all projects	1,606	91.15%	6	0.34%	0	[illegible]	0	0	2	0.11%

b. Total Math (Concluded)

Project	Absent: Number of Students	Absent: Percent of Rostered Students	Wrong Subtest: Number of Students	Wrong Subtest: Percent of Rostered Students	Special Education: Number of Students	Special Education: Percent of Rostered Students	Student Has Handicap: Number of Students	Student Has Handicap: Percent of Rostered Students	Other: Number of Students	Other: Percent of Rostered Students	Total Students Rostered
Catch-Up											
Bloomington	13	6.77%							1	0.52%	192
Brookport	5	8.20									61
Galax	5	.81									64
Providence Forge	6	9.[illegible]8									64
Wayne City	1	2.38									42
Total Catch-Up	30	7.09%							1	0.24%	423
HIT											
Lexington	15	12.30%	1	0.82%							122
Olean	8	4.0							1	0.50%	200
Total HIT	23	7.14%	1	0.31%					1	0.31%	322
R-3											
Charlotte	20	6.80%									294
Lake Village	5	3.68									136
Lorain	22	6.02			6	1.64%			6	1.64%	365
Schenectady	22	9.91			1	0.45			8	3.60	222
Total R-3	[illegible]	6.78%			7	0.69%			16	1.57%	1,017
Lorain	22	6.02			6	1.64%			6	1.64%	365
Schenectady	22	9.91			1	0.45			8	3.60	222

Note: Total Math not applicable for Conquest, IRIT, and PTR.

Table 3-6

NUMBER OF INVALID TESTS, BY REASON FOR INVALIDATION: SPRING

a. Total Reading

Project	Valid Tests		Reason for Invalidation							
			Refused Response		Borrowed Answers		Multiple Answers		Student Ill During Test	
	Number of Students	Percent of Rostered Students	Number of Students	Percent of Rostered Students	Number of Students	Percent of Rostered Students	Number of Students	Percent of Rostered Students	Number of Students	Percent of Rostered Students
Catch-Up										
Bloomington	167	86.98%								
Bridgeport	67	77.65								
Galax	49	76.56								
Providence Forge	57	89.06								
Wayne City	41	97.62								
Total Catch-Up	361	85.34%								
Conquest										
Benton Harbor	114	81.43%								
Cleveland	337	74.23	1	0.22%						
Gloversville	142	64.55								
Total Conquest	593	72.85%	1	0.12%						
HIT										
Lexington	212	87.24%							1	0.41%
Olean	206	89.96								
Total HIT	418	88.37%							1	0.21%
[illegible]										
Bloomington	41	53.95%	1	1.32%						
Oklahoma City	28	59.57					1	2.13%		
Schenectady	31	55.96								
Total [illegible]	100	56.52%	1	0.59%			1	0.59%		
PSR										
Canton	158	88.69%	4	2.12%			5	2.65%		
Dallas	114	61.96	2	1.09			5	2.72		
Total PSR	272	75.61%	6	1.61%			10	2.68%		
R-3										
Charlotte	233	79.25%								
Lake Village	124	91.18	1	0.74%						
Lorain	280	76.71								
Schenectady	175	78.83	1	0.45						
Total R-3	812	79.84%	2	0.20%						
Total, all projects	2,566	78.67%	10	0.31%	[illegible]	0	11	0.34%	1	0.03%

63

Table I-6 (Continued)

a. Total Reading (Concluded)

Project	Reason for Invalidation: Absent — Number of Students	Absent — Percent of Rostered Students	Wrong Subtest — Number of Students	Wrong Subtest — Percent of Rostered Students	Special Education — Number of Students	Special Education — Percent of Rostered Students	Absent in Fall — Number of Students	Absent in Fall — Percent of Rostered Students	Withdrawn — Number of Students	Withdrawn — Percent of Rostered Students	Total Students Rostered
Catch-Up											
Bloomington	9	4.69%			1		9	4.69%	6	3.13%	192
Brockport	5	8.20					2	3.28	7	11.48	61
Galax	1	1.56					5	7.81	9	14.06	64
Providence Forge	3	4.69					2	3.13	2	3.13	64
Wayne City							1	2.38			42
Total Catch-Up	18	4.26%			1	0.24%	19	4.49%	24	5.67%	423
Conquest											
Benton Harbor	5	3.57%					8	5.71%	13	9.29%	140
Cleveland	25	5.51					30	6.61	61	13.44	454
Gloversville	7	3.18					7	3.18	64	29.09	220
Total Conquest	37	4.55%					45	5.53%	138	16.95%	814
HIT											
Lexington	10	4.12%					5	2.06%	15	6.17%	243
Olean	8	3.49			1	0.44%	9	3.91	6	2.61	230
Total HIT	18	3.81%			1	0.21%	14	2.96%	13	2.75%	473
IRIT											
Bloomington	1	1.32%					3	3.95%	30	39.47%	76
Oklahoma City	2	4.26					2	4.26	14	29.79	47
Schenectady	2	4.26					3	6.38	11	23.40	47
Total IRIT	5	2.94%					8	4.71%	55	32.35%	170
PTR											
Canton					4	2.12%	3	1.59%	5	2.65%	189
Dallas	13	7.07%					22	11.96	28	15.22	184
Total PTR	13	3.49%			4	1.07%	25	6.70%	33	8.85%	373
R-3											
Charlotte	28	9.52%					17	5.78%	16	5.44%	294
Lake Village	3	2.21					2	1.47	6	4.41	136
Lorain	49	13.42					27	7.40	9	2.47	365
Schenectady	28	12.61					18	8.11			222
Total R-3	108	10.62%					64	6.29%	31	3.05%	1,017
Total, all projects	199	6.09%	0	0	6	0.18%	175	5.35%	294	8.99%	3,270

Table 3-6 (Concluded)

b. Total Math

Project	Valid Tests: Number of Students	Valid Tests: Percent of Rostered Students	Reason for Invalidation — Refused Response: Number of Students	Refused Response: Percent of Rostered Students	Borrowed Answers: Number of Students	Borrowed Answers: Percent of Rostered Students	Multiple Answers: Number of Students	Multiple Answers: Percent of Rostered Students	Student Ill During Test: Number of Students	Student Ill During Test: Percent of Rostered Students
Catch-Up										
Bloomington	155	80.73%	1	0.52%						
Brookport	48	78.69								
Galax	46	71.88								
Providence Forge	50	78.13	1	1.56						
Wayne City	41	97.62								
Total Catch-Up	340	80.38%	2	0.47%						
HIT										
Lexington	93	76.23%								
Olean	175	87.50								
Total HIT	268	83.23%								
R-3										
Charlotte	222	75.51%								
Lake Village	122	89.71	1	0.74%						
Lorain	295	80.82								
Schenectady	173	77.93	2	0.90						
Total R-3	812	79.84%	3	0.29%						
Total, all projects	1,420	80.59%	5	0.28%	0	0	0	0	0	0

b. Total Math (Concluded)

Project	Reason for Invalidation — Absent: Number of Students	Absent: Percent of Rostered Students	Wrong Subtest: Number of Students	Wrong Subtest: Percent of Rostered Students	Special Education: Number of Students	Special Education: Percent of Rostered Students	Absent in Fall: Number of Students	Absent in Fall: Percent of Rostered Students	Withdrawn: Number of Students	Withdrawn: Percent of Rostered Students	Total Students Rostered
Catch-Up											
Bloomington	19	9.90%			1	0.52%	13	6.77%	3	0.02%	192
Brookport	4	6.56					5	8.20	4	0.07	61
Galax	4	6.25					5	7.81	9	0.14	64
Providence Forge	7	10.94					6	9.38			64
Wayne City							1	2.38			42
Total Catch-Up	34	8.04%			1	0.24%	30	7.09%	16	0.04%	423
HIT											
Lexington	7	5.74%					15	12.30%	7	5.74%	122
Olean	10	5.00					8	4.00	7	3.50	200
Total HIT	17	5.28%					23	7.14%	14	6.83%	322
R-3											
Charlotte	38	12.93%					20	6.80%	14	4.76%	294
Lake Village	5	3.68					5	3.68	3	2.21	136
Lorain	32	8.77					22	6.02	16	4.38	365
Schenectady	28	12.61					22	9.91			222
Total R-3	103	10.13%					69	6.78%	33	3.24%	1,017
Total, all projects	154	8.74%	0	0	1	0.06%	119	6.75%	63	3.58%	1,762

Note: Total Math not applicable for Conquest, IRIT, and PTR.

3.6 Generalizability of Test Results

We do not feel that the techniques of sampling theory statistics can be applied to the data. As usual in this type of evaluation, no sampling of children, teachers, or sites was possible. We do not know how to define either the sample space or its associated probability measure.

Consequently, relative frequency procedures, Neyman-Pearson significance tests, and confidence intervals do not have their usual empirical justification. In particular, the norm-referenced analysis, because it is just such a procedure, loses its empirical justification. A further consequence of the inability to sample is the haphazard distribution of sample sizes. If we pretend that sampling theory applies, the usual inferential statistics act as if we were more interested in comparisons with larger sample sizes. For example, the norm-referenced analysis will reflect a presumed interest in R-3 and a presumed disinterest in IRIT.

Even had we been able to follow the canons of sampling theory (sampling children to be assigned to PIPs, schools, and teachers), the resulting generalizability would not have been a very useful feature of our evaluation. This is because, as a result of our work, the PIPs we evaluated will probably not be used again. In fact, one of the main purposes of this report is to justify this recommendation and to describe how we came to make it.

3.7 Achieving Criterion Growth

3.7.1 Norm-Referenced Analysis Results

The results of the norm-referenced analysis are presented in this section. Table 3-7 shows the one-third standard deviation criterion for growth. We demanded that the lower limit of the 95% confidence interval corresponding to the t test* (Eq. 2-2) be greater than the criterion before we said that the criterion was met. We also demanded that the upper limit of this confidence interval be less than the criterion before

---

* Our computer program for the analysis uses the normality assumption to interpolate between percentiles to calculate expected growth if an exact value is not found in the norm tables. The method of converting standard scores to percentile ranks and determining expected spring scores for the norm-referenced analysis is described in Appendix B. The computer program was written by SRI's George Black.

we conclude that the criterion was not met. Similar conventions were applied to the conclusions about normal growth. If the lower limit of the 95% confidence interval was greater than zero, we said that normal growth was achieved. If the upper limit was less than zero, we concluded that normal growth was not achieved. This procedure collapses the two-part RMC criterion (see page 2) into a single test.

Table 3-7

ONE-THIRD STANDARD DEVIATION OF THE MAT
NORM STANDARD SCORES FOR SPRING

Grade	Word Knowledge	Reading	Math Computation	Math Concepts	Math Problem Solving	Total Reading	Total Math
1[*]	2.9[†]	2.0					
1	2.8[‡]	3.4					4.2[§]
2	3.4	3.9	3.3	3.9	4.0	3.6	3.7
3	4.0	4.5	3.6	4.[illegible]	4.4	4.3	4.0
4	4.3	4.8	4.0	4.0	4.4	4.8	4.0
5	4.3	4.1	3.7	4.2	4.3	4.3	4.1
6	4.6	4.2	4.1	4.8	4.5	4.5	4.2
7	4.8	5.4	4.2	4.2	4.7	5.3	4.3
8	5.2	5.7	4.8	4.9	5.0	5.6	4.8
9	5.1	5.4	5.3	5.2	4.9	5.4	4.7

Note: Values in this table for grades 1-8 are derived from tables in the Metropolitan Achievement Test Special Report No. 8, Summary Statistics for National Standardization Groups (Harcourt Brace Jovanovich, Inc., New York, New York, June 1971). Values for grade 9 are computed from the Standard Score to Percentile Rank table in the Metropolitan Achievement Tests, Teacher's Handbook (Advanced) (Harcourt Brace Jovanovich, Inc., New York, New York, 1971).

[*] Primer test battery given in Canton grade 1 only.

[†] Listening for Sounds subtest.

[‡] Word Analysis subtest.

[§] Mathematics subtest.

The use of these confidence intervals does not imply that we endorse their relative frequency interpretation in the PIP study. As discussed, the conditions under which the PIP field trials were conducted make it difficult to justify applying the techniques of sampling theory to the data.

Table 3-8 shows the results of the norm-referenced analysis for PIPs, by grade and subtest. This table is most appropriate for those wishing a global view of PIP success, since projects are not distinguished here. Overall, the table shows that the PIPs did not result in projects that produced educationally significant growth. Of the 21 PIP and grade combinations that provided enough data to determine improvement in Total Reading, all showed that criterion growth was not achieved. In two instances, there was not enough information to reach a decision. Of the six PIP and grade combinations that provided enough data to decide in Total Math, four showed that criterion growth was not achieved. In nine cases, there was not enough information to reach a decision.

For grades higher than the first, the PIPs did not retard growth from the equipercentile expectation. Of the 20 Total Reading and Total Math PIP and grade combinations shown in Table 3-8 for which there were enough data to decide, we concluded that normal growth was maintained in 19. At Catch-Up grade 4, we concluded that equipercentile growth was not maintained in Total Reading.

Because the total scores are sums of the subtest scores, we were not surprised to see the same picture prevailing for the subtests. Generally, the PIPs produced projects that were more successful in producing gains on the mathematics subtests than on the reading subtests.

Table 3-8 also illustrates some points already discussed. The first grade was expected to make very large gains compared with higher grades. However, the observed gains over fall were not proportionately great. The erratic nature of the equipercentile growth curves (Figure 2-1) is apparent in that the equipercentile expected growth was negative for some PIP, grade, and subtest combinations. We see this whenever the gain over expected is greater than the gain over fall, as in HIT eighth grade Math Computations. When the gain over fall equals the gain over expected, the equipercentile growth is zero, as in R-3 eighth grade Total Reading.

Except for PTR (two sites) and Catch-Up (four to five sites), the PIPs produced projects that showed growth consistent with the equipercentile expectation. Except at the first and seventh grades, the PIP projects beat the equipercentile expectation at least as often as they failed it. However, at the first grade they failed it by wide margins in the three PIPs with a first grade.

As discussed in Section 4, our fieldwork shows the projects to have been reasonably implemented, but the six PIPs evaluated failed to produce educationally significant growth in the sense of the norm-referenced analysis. However, in the same sense, the PIP projects did generally achieve equipercentile growth.

These results do not imply that some of the projects were not successful. Tables 3-9 through 3-14 show the results of analyses of variance applied to the standard scores of projects for a given grade--for Catch-Up, Conquest, HIT, IRIT, R-3, and PTR. These tables also show the basic statistics that entered into the norm-referenced analysis shown in Table 3-8. Table 3-9 through 3-14 show that, with the exception of the Catch-Up and Conquest projects, there were project differences on Total Reading or Total Math in either the fall or spring (significant differences at $p \leq 0.05$) at all grades for all PIPs. This suggests that the global norm-referenced analyses shown in Table 3-8 were not representative of all the sites for a given PIP.

Table 3-15 shows the results of the norm-referenced analysis conducted by grade, site, and subtest. In spite of the significant F tests for project differences, the table shows fundamentally the same picture as Table 3-8. Where there were enough data to reach a decision, the decision is that criterion growth was not achieved in Total Reading. The PIP did better on the Total Math standard scores; here, if there was enough information to reach a decision, in over one-third of the cases, the decision is that criterion growth was achieved. However, the sample size becomes so small for many project and grade combinations that we were not able to reach a decision over half the time.

The result drawn from Table 3-8 concerning the achievement of normal growth is also substantiated at the project level; in the 22 cases where we could reach a decision on Total Reading using the norm-referenced procedure, 19 cases confirmed normal growth. The three exceptions were all in the fourth grade, one of the two grades in which we gave the same level of the MAT in fall and spring. In the 18 cases where a decision could be reached on normal growth in Total Math, all 18 were favorable.

Also consistent with Table 3-8 is that the PIP projects were able to meet the equipercentile growth expectation, often showing large (although generally nonsignificant) gains over expected. For example, the IRIT third grade projects all showed about 5-point gains at grade 3 for the Reading subtest. In Math Computation, as well as in Reading, Brookport and Galax Catch-Up made similar large (but nonsignificant) gains over expected in both the fifth and sixth grades.

Table 3-8

RESULTS OF THE NORM-REFERENCED ANALYSIS, BY PIP, GRADE, AND SUBTEST

PIP	Word Knowledge					Reading				
	Gain over Fall	Gain over Expected Gain	t Test	Meets Criterion Growth	Meets Normal Growth	Gain over Fall	Gain over Expected Gain	t Test	Meets Criterion Growth	Meets Normal Growth
Catch-Up										
Grade 1										
4 sites	7.32	-4.93	-6.14	No*	No*	10.28	-4.22	-4.60	No	No
No. pupils			47					47		
Grade 2										
4 sites	12.50	1.50	1.67	No	Unknown	5.92	-1.78	-1.85	No	Unknown
No. pupils			50					50		
Grade 3										
4 sites	3.14	0.14	0.16	No	Unknown	3.63	1.67	0.89	Unknown	Unknown
No. pupils			51					40		
Grade 4										
5 sites	3.17	-0.83	-1.34	No	Unknown	1.76	-3.07	-3.79	No	No
No. pupils			83					83		
Grade 5										
5 sites	3.88	-0.37	-0.51	No	Unknown	9.74	3.74	3.52	Unknown	Yes
No. pupils			73					73		
Grade 6										
4 sites	2.18	-0.82	-0.99	No	Unknown	5.88	0.87	0.68	No	Unknown
No. pupils			39					40		
Grade 7										
1 site	1.20	0.40	0.37	Unknown	Unknown	-0.20	-0.60	-0.30	No	Unknown
No. pupils			5					5		
Grade 8										
1 site	6.75	5.00	2.09	Unknown	Unknown	5.00	3.12	2.89	Unknown	Unknown
No. pupils			4					4		
Conquest										
Grade 1										
1 site	4.86	-14.90	-15.06	No*	No*	10.80	-13.70	-13.50	No	No
No. pupils			21					20		
Grade 2										
3 sites	10.14	-0.38	-0.74	No	Unknown	9.00	1.00	1.79	No	Unknown
No. pupils			169					163		
Grade 3										
3 sites	3.90	0.90	1.60	No	Unknown	5.72	3.72	5.08	Unknown	Yes
No. pupils			115					138		
Grade 4										
3 sites	5.07	0.07	0.13	No	Unknown	4.76	-0.09	-0.12	No	Unknown
No. pupils			106					111		
Grade 5										
3 sites	6.50	1.50	2.42	No	Yes	10.55	4.55	4.50	Unknown	Yes
No. pupils			68					69		
Grade 6										
3 sites	5.37	2.37	3.67	No	Yes	7.56	1.56	1.74	No	Unknown
No. pupils			87					87		

PIP	Math Computation					Math Concepts					Math Problem Solving				
	Gain over Fall	Gain over Expected Gain	t Test	Meets Criterion Growth	Meets Normal Growth	Gain over Fall	Gain over Expected Gain	t Test	Meets Criterion Growth	Meets Normal Growth	Gain over Fall	Gain over Expected Gain	t Test	Meets Criterion Growth	Meets Normal Growth
Catch-Up															
Grade 1															
4 sites															
No. pupils															
Grade 2															
4 sites															
No. pupils															
Grade 3															
4 sites	9.20	1.20	1.07	No	Unknown	3.62	-2.99	-2.56	No	No	4.85	-0.15	-0.11	No	Unknown
No. pupils			46					45					52		
Grade 4															
5 sites	7.91	1.59	1.69	No	Unknown	5.73	-0.03	-0.03	No	Unknown	5.73	-0.27	-0.24	No	Unknown
No. pupils			81					81					77		
Grade 5															
5 sites	11.85	7.85	7.57	Yes	Yes	9.36	6.91	6.04	Yes	Yes	8.00	4.44	3.89	Unknown	Yes
No. pupils			71					70					70		
Grade 6															
4 sites	11.80	6.80	5.08	Unknown	Yes	8.49	5.49	4.44	Unknown	Yes	7.42	3.45	3.06	Unknown	Yes
No. pupils			40					39					38		
Grade 7															
1 site	5.20	2.30	0.89	Unknown	Unknown	-4.20	-4.70	-0.95	Unknown	Unknown	10.60	8.60	1.28	Unknown	Unknown
No. pupils			5					5					5		
Grade 8															
1 site	5.25	3.37	1.15	Unknown	Unknown	5.00	5.00	2.40	Unknown	Unknown	2.00	0.00	0.00	Unknown	Unknown
No. pupils			4					4					4		

PIP	Total Reading					Total Math				
	Gain over Fall	Gain over Expected Gain	t Test	Meets Criterion Growth	Meets Normal Growth	Gain over Fall	Gain over Expected Gain	t Test	Meets Criterion Growth	Meets Normal Growth
Catch-Up										
Grade 1										
4 sites						10.26	-10.67	-11.05	No*	No*
No. pupils								46		
Grade 2										
4 sites	9.52	-0.41	-0.58	No	Unknown	16.59	1.59	1.35	Unknown	Unknown
No. pupils			50					46		
Grade 3										
4 sites	2.70	0.31	0.25	No	Unknown	6.30	-1.70	-1.63	No	Unknown
No. pupils			40					44		
Grade 4										
5 sites	2.96	-2.04	-3.46	No	No	7.64	0.64	0.85	No	Unknown
No. pupils			83					75		
Grade 5										
5 sites	6.66	1.66	2.34	No	Yes	11.25	5.62	5.7.	Unknown	Yes
No. pupils			73					65		
Grade 6										
4 sites	3.97	-0.03	-0.03	No	Unknown	10.55	7.36	8.09	Yes	Yes
No. pupils			39					38		
Grade 7										
1 site	0.00	-0.00	-0.00	No	Unknown	4.60	3.10	0.93	Unknown	Unknown
No. pupils			5					5		
Grade 8										
1 site	6.75	5.75	2.64	Unknown	Unknown	5.00	5.00	4.63	Unknown	Yes
No. pupils			4					4		
Conquest										
Grade 1										
1 site										
No. pupils										
Grade 2										
3 sites	9.87	0.80	2.09	No	Yes					
No. pupils			159							
Grade 3										
3 sites	3.85	1.70	2.78	No	Yes					
No. pupils			115							
Grade 4										
3 sites	5.56	1.08	1.98	No	Unknown					
No. pupils			108							
Grade 5										
3 sites	8.69	3.69	5.45	Unknown	Yes					
No. pupils			68							
Grade 6										
3 sites	6.95	2.95	4.84	No	Yes					
No. pupils			87							

Table 3-8 (Concluded)

PIP	Word Knowledge: Gain over Fall	Word Knowledge: Gain over Expected Gain	Word Knowledge: t Test	Word Knowledge: Meets Criterion Growth	Word Knowledge: Meets Normal Growth	Reading: Gain over Fall	Reading: Gain over Expected Gain	Reading: t Test	Reading: Meets Criterion Growth	Reading: Meets Normal Growth
HIT										
Grade 6										
1 site										
Tutors	5.73	1.73	1.69	No	Unknown	3.68	-1.32	-0.99	No	Unknown
No. pupils			22					22		
Tutees	4.12	0.05	0.05	No	Unknown	7.12	0.19	0.14	No	Unknown
No. pupils			41					41		
Grade 7										
2 sites										
Tutors	1.99	0.27	0.42	No	Unknown	1.48	0.98	1.25	No	Unknown
No. pupils			90					90		
Tutees	0.66	0.66	0.69	No	Unknown	0.25	-0.75	-0.66	No	Unknown
No. pupils			64					64		
Grade 8										
2 sites										
Tutors	3.41	1.41	1.61	No	Unknown	3.15	3.15	2.75	No	Yes
No. pupils			46					46		
Tutees	2.84	0.87	0.98	No	Unknown	3.00	1.00	0.91	No	Unknown
No. pupils			63					63		
Grade 9										
2 sites										
Tutors	0.69	-1.31	-1.85	No	Unknown	3.08	1.51	1.46	No	Unknown
No. pupils			71					72		
Tutees	3.72	1.72	1.96	No	Unknown	1.67	-2.36	-1.07	No	Unknown
No. pupils			18					18		
IRIT										
Grade 3										
3 sites	4.94	1.44	2.28	No	Yes	6.61	5.04	5.52	Unknown	Yes
No. pupils			67					67		
Grade 4										
2 sites	3.79	-1.21	-1.05	No	Unknown	2.79	-1.21	-0.80	No	Unknown
No. pupils			34					34		
R-3										
Grade 8										
4 sites	3.75	1.75	7.06	No	Yes	4.28	4.28	13.60	No	Yes
No. pupils			780					790		

PIP	Math Computation: Gain over Fall	Math Computation: Gain over Expected Gain	Math Computation: t Test	Math Computation: Meets Criterion Growth	Math Computation: Meets Normal Growth	Math Concepts: Gain over Fall	Math Concepts: Gain over Expected Gain	Math Concepts: t Test	Math Concepts: Meets Criterion Growth	Math Concepts: Meets Normal Growth
HIT										
Grade 6										
1 site										
Tutors	8.36	3.90	1.70	Unknown	Unknown	5.09	2.09	1.29	Unknown	Unknown
No. pupils			11					11		
Tutees	12.92	10.92	7.52	Yes	Yes	9.35	10.45	7.68	Yes	Yes
No. pupils			25					26		
Grade 7										
2 sites										
Tutors	4.45	1.95	2.41	No	Yes	0.72	-0.84	-0.91	No	Unknown
No. pupils			74					74		
Tutees	5.40	4.05	3.62	Unknown	Yes	1.62	1.01	0.89	No	Unknown
No. pupils			50					50		
Grade 8										
2 sites										
Tutors	0.21	1.16	0.45	Unknown	Unknown	-2.26	-2.26	-1.30	No	Unknown
No. pupils			19					19		
Tutees	3.69	4.69	3.89	Unknown	Yes	1.91	1.91	1.80	No	Unknown
No. pupils			45					46		
Grade 9										
2 sites										
Tutors	4.33	2.65	2.52	No	Yes	3.50	3.50	3.33	Unknown	Yes
No. pupils			45					46		
Tutees										
No. pupils										
IRIT										
Grade 3										
3 sites										
No. pupils										
Grade 4										
2 sites										
No. pupils										
R-3										
Grade 8										
4 sites	2.86	3.86	14.83	No	Yes	3.35	3.35	12.39	No	Yes
No. pupils			786					780		

PIP	Math Problem Solving: Gain over Fall	Math Problem Solving: Gain over Expected Gain	Math Problem Solving: t Test	Math Problem Solving: Meets Criterion Growth	Math Problem Solving: Meets Normal Growth	Total Reading: Gain over Fall	Total Reading: Gain over Expected Gain	Total Reading: t Test	Total Reading: Meets Criterion Growth	Total Reading: Meets Normal Growth
HIT										
Grade 6										
1 site										
Tutors	6.10	3.10	0.90	Unknown	Unknown	5.59	1.59	1.86	No	Unknown
No. pupils			10					22		
Tutees	7.60	4.60	4.14	Unknown	Yes	5.71	1.29	1.38	No	Unknown
No. pupils			25					41		
Grade 7										
2 sites										
Tutors	8.77	6.55	8.86	Yes	Yes	1.83	0.88	1.55	No	Unknown
No. pupils			73					90		
Tutees	6.78	4.51	3.27	Unknown	Yes	0.33	-0.18	-0.22	No	Unknown
No. pupils			50					64		
Grade 8										
2 sites										
Tutors	2.63	1.71	0.84	Unknown	Unknown	3.74	3.07	3.86	No	Yes
No. pupils			19					46		
Tutees	2.84	1.84	1.86	No	Unknown	3.24	2.24	7.83	No	Yes
No. pupils			44					62		
Grade 9										
2 sites										
Tutors	4.41	2.61	2.42	No	Yes	1.87	1.87	2.51	No	Yes
No. pupils			44					71		
Tutees						3.17	1.17	0.98	No	Unknown
No. pupils								18		
IRIT										
Grade 3										
3 sites						5.14	2.42	3.63	No	Yes
No. pupils								66		
Grade 4										
2 sites						3.44	-1.23	-1.39	No	Unknown
No. pupils								34		
R-3										
Grade 8										
4 sites	3.51	2.58	8.39	No	Yes	4.38	3.58	16.03	No	Yes
No. pupils			782					779		

PIP	Total Math: Gain over Fall	Total Math: Gain over Expected Gain	Total Math: t Test	Total Math: Meets Criterion Growth	Total Math: Meets Normal Growth
HIT					
Grade 6					
1 site					
Tutors	6.90	2.90	1.46	Unknown	Unknown
No. pupils			10		
Tutees	11.04	7.04	6.05	Yes	Yes
No. pupils			23		
Grade 7					
2 sites					
Tutors	5.68	3.68	6.15	Unknown	Yes
No. pupils			72		
Tutees	5.17	3.17	3.58	Unknown	Yes
No. pupils			48		
Grade 8					
2 sites					
Tutors	-0.37	-0.37	-0.18	No	Unknown
No. pupils			19		
Tutees	2.75	2.75	2.56	Unknown	Yes
No. pupils			40		
Grade 9					
2 sites					
Tutors	4.08	4.08	6.11	Unknown	Yes
No. pupils			40		
Tutees					
No. pupils					
IRIT					
Grade 3					
3 sites					
No. pupils					
Grade 4					
2 sites					
No. pupils					
R-3					
Grade 8					
4 sites	3.38	3.38	15.96	No	Yes
No. pupils			774		

Note: See Table 3-15 for PTR.

Primer--Listening for Sounds; Primary I--Word Analysis.

Primer--Numbers; Primary I--Mathematics.

Table 3-9

DESCRIPTIVE STATISTICS FOR CATCH-UP, BY GRADE

	Statistic	Word Knowledge				Reading				Math Computation			
		Fall	Exp	Spring	t Test	Fall	Exp	Spring	t Test	Fall	Exp	Spring	t Test
Grade 2 (4 sites)	Mean	34.55	46.58	48.08	1.67	35.38	43.40	41.62	-1.85	33.07	47.06	48.92	1.93
Primary I, fall	SD	7.70		4.92		4.23		7.75		4.05		6.08	
Primary II, spring	N	55	(47)	50		55	(50)	50		54	(48)	49	
	BMS	219.63		80.33		8.77		125.11		49.84		91.26	
	WMS	49.90		20.54		18.44		55.75		14.36		33.33	
	F ratio	4.40*		3.91*		0.48		2.24		3.47*		2.74*	
	F-S			0.50				0.48				0.14	
Grade 3 (4 sites)	Mean	49.54	52.71	52.96	0.16	44.02	47.03	49.23	0.89	48.38	56.28	57.35	1.07
Primary II, fall	SD	6.41		7.49		7.80		10.34		7.79		9.33	
Elementary, spring	N	59	(51)	52		47	(40)	52		58	(46)	49	
	BMS	142.09		165.76		83.58		47.61		235.96		217.94	
	WMS	35.57		49.26		59.26		110.59		50.92		74.29	
	F ratio	3.99*		3.36*		1.41		0.43		4.63*		3.74*	
	F-S			0.62				0.16				0.58	
Grade 4 (5 sites)	Mean	55.52	59.48	58.65	-1.34	53.41	58.17	55.10	-3.79	56.84	63.41	65.00	1.69
Elementary, fall	SD	7.95		8.08		7.68		9.06		8.11		10.53	
Elementary, spring	N	93	(83)	83		93	(83)	83		91	(81)	81	
	BMS	98.59		122.67		110.61		93.17		241.78		572.84	
	WMS	61.58		62.41		56.66		81.49		57.50		86.53	
	F ratio	1.60		1.97		1.95		1.14		4.20*		6.62*	
	F-S			0.75				0.63				0.62	
Grade 5 (5 sites)	Mean	61.20	65.75	65.27	-0.51	56.95	63.21	67.12	3.52	62.47	66.92	74.53	7.37
Elementary, fall	SD	7.56		7.48		9.64		10.48		10.15		8.79	
Intermediate, spring	N	79	(73)	77		79	(73)	77		81	(71)	73	
	BMS	84.58		62.20		218.90		93.25		394.69		396.66	
	WMS	55.68		55.65		86.03		110.85		87.65		58.49	
	F ratio	1.52		1.12		2.54*		0.84		4.50*		6.78*	
	F-S			0.67				0.59				0.58	
Grade 6 (4 sites)	Mean	66.11	68.72	67.81	-0.99	62.87	67.98	68.86	0.68	69.52	75.03	81.36	5.08
Elementary, fall	SD	8.02		9.83		10.66		11.47		10.59		9.26	
Intermediate, spring	N	46	(39)	42		47	(40)	42		44	(40)	42	
	BMS	116.60		115.39		290.94		417.34		313.88		348.15	
	WMS	60.54		95.11		101.17		109.08		96.98		64.98	
	F ratio	1.93		1.21		2.88*		3.83*		3.24*		5.36*	
	F-S			0.85				0.75				0.63	
Grade 7 (1 site)	Mean	78.40	79.20	79.60	0.17	82.70	82.60	52.00	-0.30	91.20	94.10	96.40	0.89
Intermediate, fall	SD	6.39		10.45		12.44		13.17		6.14		7.50	
Advanced, spring	N	5	(5)	5		5	(5)	5		5	(5)	5	
	F-S			0.91				0.94				0.65	
Grade 8 (1 site)	Mean												
Advanced, fall	SD												
Advanced, spring	N	4											

	Statistic	Math Concepts				Problem Solving				Total Reading				Total Math			
		Fall	Exp	Spring	t Test	Fall	Exp	Spring	t Test	Fall	Exp	Spring	t Test	Fall	Exp	Spring	t Test
Grade 2 (4 sites)	Mean			46.18				46.73		34.42	45.07	44.66	-0.58	33.07	48.04	49.57	1.35
Primary I, fall	SD			8.53				9.19		5.38		5.49		4.05		7.46	
Primary II, spring	N			49				48		55	(50)	50		54	(46)	47	
	BMS			112.48				149.03		85.36		92.64		49.84		116.33	
	WMS			70.18				79.96		25.67		26.11		14.36		51.36	
	F ratio			1.60				1.86		2.33*		3.55*		3.47*		2.27	
	F-S											0.55				0.13	
Grade 3 (4 sites)	Mean	50.63	58.37	55.00	-2.56	49.08	54.02	53.67	-0.11	45.96	48.81	49.79	0.25	51.33	60.34	58.23	-1.63
Primary II, fall	SD	9.08		9.19		8.63		11.77		5.92		8.46		7.50		9.52	
Elementary, spring	N	57	(45)	49		62	(52)	52		47	(40)	57		57	(44)	48	
	BMS	315.61		198.85		179.19		60.79		75.20		120.49		208.03		177.12	
	WMS	69.18		76.74		68.98		143.41		32.24		68.57		47.71		84.66	
	F ratio	4.56*		2.59		2.60		0.42		2.33		1.76		4.36*		2.09	
	F-S			0.62				0.53				0.49				0.65	
Grade 4 (5 sites)	Mean	54.28	60.24	60.21	-0.03	54.78	61.12	60.80	-0.24	53.03	57.95	55.92	-3.46	57.95	65.01	65.54	0.85
Elementary, fall	SD	8.65		8.42		9.44		9.89		7.39		8.19		8.09		8.74	
Elementary, spring	N	90	(81)	81		89	(77)	80		93	(83)	83		87	(75)	78	
	BMS	45.63		32.92		121.25		135.44		119.01		[illegible]		181.10		202.67	
	WMS	76.11		72.94		87.63		95.91		51.62		65.42		59.80		64.52	
	F ratio	0.60		0.45		1.38		1.41		2.31		1.53		3.03*		2.92*	
	F-S			0.54				0.48				[illegible]				0.70	
Grade 5 (5 sites)	Mean	58.65	61.10	67.79	6.04	60.90	64.44	69.03	3.89	56.30	63.59	65.32	2.34	63.90	69.75	75.15	5.70
Elementary, fall	SD	10.58		8.44		9.87		8.66		8.07		8.87		9.54		8.08	
Intermediate, spring	N	81	(70)	72		81	(70)	71		79	(73)	77		79	(65)	68	
	BMS	615.00		180.07		333.26		[illegible]		134.21		79.82		494.37		207.91	
	WMS	85.37		64.74		85.03		70.39		61.40		78.66		59.24		56.10	
	F ratio	7.20*		2.78*		3.92*		2.16		2.19		1.01		7.14*		3.70*	
	F-S			0.53				0.49				0.75				0.60	
Grade 6 (4 sites)	Mean	65.34	68.21	74.05	4.44	67.33	72.03	75.10	3.06	63.89	67.79	67.71	-0.03	71.09	74.80	82.33	8.09
Elementary, fall	SD	9.99		8.57		10.67		9.63		9.05		10.65		9.58		7.21	
Intermediate, spring	N	44	(39)	40		45	(38)	40		46	(39)	42		44	(39)	39	
	BMS	190.03		254.92		558.98		[illegible]		217.17		248.57		296.86		191.37	
	WMS	93.10		58.36		81.39		81.11		72.67		102.71		76.33		39.96	
	F ratio	2.04		4.37*		6.88*		2.85		2.92*		2.42		3.59*		4.79*	
	F-S			0.68				0.79				0.87				0.83	
Grade 7 (1 site)	Mean	93.40	93.90	89.20	-0.95	86.60	88.60	97.20	1.28	80.40	80.40	80.40	0	94.20	95.70	98.30	0.93
Intermediate, fall	SD	13.72		9.07		14.29		8.11		9.10		12.62		9.76		6.67	
Advanced, spring	N	5	(5)	5		5	(5)	5		5	(5)	5		5	(5)	5	
	F-S			0.59				0.19				1.00				0.65	
Grade 8 (1 site)	Mean																
Advanced, fall	SD																
Advanced, spring	N																

Key: Exp is the expected mean; SD = standard deviation; N = number of students (number in parentheses are the numbers of children for whom both fall and spring data are available); BMS = between mean square; WMS = within mean square; F-S is the fall to spring correlation.

*p = ≤ 0.05.

Table 3-10

DESCRIPTIVE STATISTICS FOR R-3, BY GRADE

R-3	Statistic	Word Knowledge Fall	Word Knowledge Exp	Word Knowledge Spring	Word Knowledge t Test	Reading Fall	Reading Exp	Reading Spring	Reading t Test	Math Computation Fall	Math Computation Exp	Math Computation Spring	Math Computation t Test	Math Concepts Fall	Math Concepts Exp	Math Concepts Spring	Math Concepts t Test
Grade 8 (4 sites)	Mean	81.01	83.32	84.89	7.06	80.84	81.20	85.14	13.80	90.47	89.98	93.67	14.83	83.22	83.78	86.84	12.39
Advanced, fall	SD	13.76		14.24		14.75		15.76		12.33		13.18		12.71		14.29	
Advanced, spring	N	929	(780)	813		932	(790)	822		933	(786)	816		930	(780)	812	
	BMS	2860.46		2624.17		1336.53		2427.31		591.45		447.94		797.20		853.97	
	WMS	180.72		193.71		213.94		240.34		150.72		172.72		159.50		201.86	
	F ratio	15.83*		13.55*		6.25*		10.10*		3.92*		2.59		5.00*		4.23*	
	F-S			0.88				0.84				0.84				0.85	

R-3	Statistic	Problem Solving Fall	Problem Solving Exp	Problem Solving Spring	Problem Solving t Test	Total Reading Fall	Total Reading Exp	Total Reading Spring	Total Reading t Test	Total Math Fall	Total Math Exp	Total Math Spring	Total Math t Test
Grade 8 (4 sites)	Mean	87.63	88.85	91.17	8.39	81.02	82.21	85.51	16.03	91.93	92.43	95.59	15.96
Advanced, fall	SD	14.52		14.77		14.66		15.51		13.10		13.91	
Advanced, spring	N	928	(782)	817		929	(779)	812		924	(774)	812	
	BMS	191.63		813.06		2463.69		2914.24		363.69		748.31	
	WMS	210.84		216.10		207.76		230.48		170.88		191.41	
	F ratio	0.91		3.76*		11.86*		17.64*		2.13		3.91*	
	F-S			0.83				0.92				0.91	

Key: Exp is the expected mean; SD = standard deviation; N = number of students (numbers in parentheses are the numbers of children for whom both fall and spring data are available); BMS = between mean square; WMS = within mean square; F-S is the fall to spring correlation.

*p = ≤0.05.

Table 3-11

DESCRIPTIVE STATISTICS FOR CONQUEST, BY GRADE

		Word Knowledge				Reading				Total Reading			
Conquest Grade	Statistic	Fall	Exp	Spring	t Test	Fall	Exp	Spring	t Test	Fall	Exp	Spring	t Test
Grade 2 (3 sites)	Mean	38.79	49.50	49.19	-0.74	36.59	44.72	45.54	1.79	36.99	46.28	46.89	2.09
Primary I, fall	SD	8.60		5.44		7.10		7.56		6.81		5.44	
Primary II, spring	N	208	(169)	170		198	(163)	173		196	(159)	170	
	BMS	631.09		33.67		383.73		233.33		524.04		64.57	
	WMS	68.46		29.55		46.99		55.04		41.40		29.21	
	F ratio	9.22*		1.14		8.17*		4.24*		12.66*		2.21	
	F-S			0.63				0.51				0.69	
Grade 3 (3 sites)	Mean	48.40	51.46	52.70	1.60	44.38	46.35	50.07	5.08	45.77	47.86	49.89	2.78
Primary II, fall	SD	6.19		7.15		7.48		8.57		5.85		7.77	
Elementary, spring	N	154	(115)	136		177	(138)	138		154	(115)	136	
	BMS	34.17		55.43		107.36		73.70		55.56		89.87	
	WMS	38.40		51.08		55.41		73.36		33.88		59.89	
	F ratio	0.89		1.09		1.94		1.00		1.64		1.50	
	F-S			0.59				0.42				0.54	
Grade 4 (3 sites)	Mean	52.51	56.71	56.67	0.13	50.28	55.15	55.05	-0.12	50.02	53.99	54.79	1.98
Elementary, fall	SD	7.29		7.58		7.89		10.55		7.28		8.04	
Elementary, spring	N	149	(108)	111		152	(111)	111		149	(108)	111	
	BMS	49.38		19.92		283.82		314.41		110.25		89.49	
	WMS	53.14		58.14		59.35		107.45		52.22		64.25	
	F ratio	0.93		0.34		4.78*		2.93		2.11		1.39	
	F-S			0.71				0.64				0.73	
Grade 5 (3 sites)	Mean	59.19	64.13	65.55	2.42	56.05	62.36	66.91	4.50	56.71	61.81	65.41	5.45
Elementary, fall	SD	7.[illegible]		6.79		9.72		8.71		7.77		7.57	
Intermediate, spring	N	94	(58)	69		96	(69)	69		94	(68)	69	
	BMS	71[illegible]		403.96		835.71		257.17		663.06		417.57	
	WMS	[illegible].52		35.20		78.59		70.35		47.20		46.42	
	F ratio	[illegible].36*		11.48*		10.63*		3.66*		14.05*		9.00*	
	F-S			0.76				0.55				0.72	
Grade 6 (3 sites)	Mean	65.27	68.70	71.07	3.67	63.83	70.77	72.33	1.74	63.79	68.45	71.40	4.84
Elementary, fall	SD	9.82		8.01		11.97		10.81		10.74		9.26	
Intermediate, spring	N	106	(87)	87		106	(87)	87		106	(87)	87	
	BMS	1983.80		1112.17		3096.08		1289.55		2608.74		1405.60	
	WMS	59.74		39.25		85.89		89.03		66.83		54.35	
	F ratio	33.21*		28.33*		36.05*		14.49*		39.03*		25.86*	
	F-S			0.78				0.72				0.84	

Key: Exp is the expected mean; SD = standard deviation; N = number of students (numbers in parentheses are the numbers of children for whom both fall and spring data are available); BMS = between mean square; WMS = within mean square; F-S is the fall to spring correlation.

*p = ≤0.05.

Table 3-12

DESCRIPTIVE STATISTICS FOR IRIT, BY GRADE

IRIT Grade	Statistic	Word Knowledge				Reading				Total Reading			
		Fall	Exp	Spring	t Test	Fall	Exp	Spring	t Test	Fall	Exp	Spring	t Test
Grade 3 (3 sites)	Mean	51.25	54.02	55.46	2.28	46.60	47.23	52.27	5.52	48.41	50.31	52.73	3.63
Primary II, fall	SD	5.86		5.70		8.47		7.42		6.44		5.98	
Elementary, spring	N	105	(67)	67		105	(67)	67		105	(65)	66	
	BMS	245.59		62.48		414.45		249.15		293.30		112.53	
	WMS	30.24		31.56		65.10		49.04		36.54		33.27	
	F ratio	8.12*		1.98		6.37*		5.08*		8.03*		3.38*	
	F-S			0.58				0.49				0.58	
Grade 4 (2 sites)	Mean	58.26	63.68	62.47	-1.05	55.19	60.74	59.53	-0.80	55.75	61.34	60.12	-1.39
Elementary, fall	SD	7.80		7.79		9.06		9.95		7.67		7.93	
Elementary, spring	N	58	(34)	34		57	(34)	34		57	(34)	34	
	BMS	84.00		543 53		4.47		450.42		42.20		480.10	
	WMS	60.38		45.53		83.50		88.06		59.10		49.92	
	F ratio	1.39		11.94*		0.05		5.11*		0.71		9.62*	
	F-S			0.70				0.61				0.81	

Key: Exp is the expected mean; SD = standard deviation; N = number of students (number in parentheses are the numbers of children for whom both fall and spring data are available); BMS = between mean square; WMS = within mean square; F-S is the fall to spring correlation.

*p = ≤ 0.05.

Table 3-13

DESCRIPTIVE STATISTICS FOR HIT, BY GRADE

HIT Grade	Statistic	Word Knowledge				Reading				Math Computation				Math Concepts				Problem Solving				Total Reading				Total Math			
		Fall	Exp	Spring	t Test	Fall	Exp	Spring	t Test	Fall	Exp	Spring	t Test	Fall	Exp	Spring	t Test	Fall	Exp	Spring	t Test	Fall	Exp	Spring	t Test	Fall	Exp	Spring	t Test
Grade 6: Tutors (1 site)	Mean	69.29	73.55	75.27	1.69	72.58	77.27	75.95	-0.99	76.33	84.55	88.45	1.70	8.60	73.27	75.36	1.29	77.69	82.00	83.27	0.90	69.96	73.95	75.55	1.86	79.46	84.80	86.64	1.46
Elementary, fall	SD	6.09		8.00		7.47		9.02		13.79		11.93		13.89		11.27		15.69		11.63		6.93		7.88		14.60		10.98	
Intermediate, spring	N	24	(22)	22		24	(22)	22		15	(11)	11		15	(11)	11		13	(10)	11		24	(22)	22		13	(10)	11	
	F-S			0.80				0.73				0.83				0.94				0.82				0.86				0.98	
Grade 6: Tutees (1 site)	Mean	58.23	62.15	62.20	0.05	58.47	65.03	65.22	0.14	59.54	62.08	72.70	7.52	55.55	54.24	64.96	7.68	59.00	62.04	66.48	4.14	57.09	61.24	62.54	1.38	61.48	65.65	72.22	6.05
Elementary, fall	SD	8.58		8.63		9.35		9.00		10.26		8.86		8.61		7.18		9.44		8.66		8.94		8.81		9.41		7.52	
Intermediate, spring	N	43	(41)	41		43	(41)	41		28	(25)	27		29	(26)	27		27	(25)	27		43	(41)	41		25	(23)	27	
	F-S			0.74				0.57				0.73				0.66				0.83				0.7[illegible]				0.82	
Grade 7: Tutors (2 sites)	Mean	86.62	88.28	88.54	0.42	86.06	86.71	87.69	1.25	98.63	101.69	103.64	2.41	93.61	96.11	95.27	-0.91	91.36	93.66	100.21	8.86	86.89	87.94	88.82	1.55	99.12	101.71	105.39	6.15
Intermediate, fall	SD	11.90		11.99		10.87		13.96		11.99		10.43		14.27		12.55		11.03		11.08		11.42		13.44		12.34		11.11	
Advanced, spring	N	93	(90)	90		93	(90)	90		76	(74)	74		76	(74)	74		76	(73)	73		93	(90)	90		76	(72)	72	
	BMS	2523.26		2579.40		1078.02		2688.68		4292.12		1782.77		4806.96		2977.90		2890.62		2051.37		1889.35		3051.16		3930.77		2195.84	
	WMS	115.43		115.97		107.62		166.58		87.62		85.48		141.42		118.45		84.28		95.59		111.14		147.95		101 12		93.82	
	F-ratio	24.86*		22.24*		10.02*		16.14*		48.99*		20.86*		33.99*		25.14*		34.30*		21.46*		17.00*		20.62*		38.87*		23.41*	
	F-S			0.87				0.85				0.81				0.81				0.84				.92				0.90	
Grade 7: Tutees (2 sites)	Mean	72.45	73.19	73.84	0.69	72.49	74.17	73.42	-0.66	84.04	85.71	89.88	3.62	76.65	77.39	78.47	0.89	77.38	79.73	84.15	3.27	72.19	73.49	73.31	-0.22	83.64	85.85	89.06	3.58
Intermediate, fall	SD	9.51		10.75		9.62		10.34		9.90		9.64		11.06		9.86		10.25		11.02		9.55		10.34		9.45		9.13	
Advanced, spring	N	67	(64)	64		67	(64)	64		52	(50)	51		52	(50)	51		52	(50)	52		67	(64)	64		50	(48)	51	
	BMS	1998.68		2539.80		1285.77		676.38		1913.86		484 76		1134.43		868.58		1400.98		2207.84		1939.07		1962.04		1662.30		1081.46	
	WMS	61.17		76.49		74.11		97.70		61.60		84.99		102.03		81.47		79.19		79.70		62.76		77.03		56.44		63.01	
	F-ratio	32.68*		33.20*		17.35*		6.92*		31.07*		5.70*		11.12*		10.66*		17.69*		27.70*		30.90*		25.47*		29.45*		17.16*	
	F-S			0.72				0.58				0.66				0.72				0.59				0.77				0.79	
Grade 8: Tutors (2 sites)	Mean	82.40	84.59	86.00	1.61	79.65	80.54	83.70	2.75	90.70	90.10	91.26	0.45	79.71	81.63	79.37	-1.30	84.85	86.08	87.79	0.84	81.15	82.34	85.41	3.86	90.55	91.26	90.89	-0.18
Advanced, fall	SD	13.93		14.31		15.04		14.66		7.79		11.21		9.59		12.30		8.07		11.13		15.06		15.31		7.71		11.78	
Advanced, spring	N	52	(46)	46		52	(46)	46		20	(19)	19		21	(19)	19		20	(19)	19		52	(46)	46		20	(19)	19	
	BMS	5733.54		5142.50		5948.17		5077.98		215.25		370.57		619.29		1340.14		470.87		670.88		6579.70		5880.14		440.21		833.79	
	WMS	83.34		92.44		111.91		104.36		52.16		111.24		64.26		81.43		42.65		91.66		99.82		105.98		38.26		97.88	
	F-ratio	68.80*		55.63*		53.15*		48.66*		4.13		3.33		9.64*		16.46*		11.04*		7.32*		65.91*		55.48*		11.50*		8.52*	
	F-S			0.91				0.86				0.35				0.81				0.61				0.94				0.66	
Grade 8: Tutees (2 sites,	Mean	76.76	78.94	79.57	0.98	73.48	75.46	76.20	0.91	90.70	88.73	93.76	3.89	81.85	81.41	83.66	1.80	89.54	90.05	92.13	1.86	74.90	76.08	78.06	2.81	92.04	91.05	94.40	2.56
reading; 1 site, math)	SD	10.29		12.45		10.61		11.06		11.27		10.52		9.20		12.29		9.34		11.05		10.39		12.12		10.13		11.64	
Advanced, fall	N	72	(63)	65		71	(63)	66		53	(45)	50		54	(46)	50		57	(44)	45		71	(62)	65		52	(40)	45	
Advanced, spring	BMS	2744.26		3468.74		2167.95		1466.27														3065.78		2877.01					
	WMS	68.15		102.43		82.72		101.35														65.14		103.47					
	F-ratio	40.27*		33.86*		26.21*		14.47*														47.07*		27.80*					
	F-S			0.82				0.67				0.75				0.82				0.81				0.85				0.83	
Grade 9: Tutors (2 sites)	Mean	86.40	89.58	88.57	-1.85	83.09	85.44	86.94	1.46	100.96	105.06	106.58	2.5.	89.59	91.35	94.85	3.33	97.50	101.61	103.81	2.42	85.32	86.41	88.63	2.51	101.48	104.02	107.60	6.11
Advanced, fall	SD	13.20		13.99		15.94		15.58		12.33		12.86		13.23		12.72		13.40		13.46		14.92		15.19		12.52		12.27	
Advanced, spring	N	85	(71)	72		86	(72)	72		55	(45)	48		58	(46)	46		54	(44)	47		85	(71)	72		54	(40)	43	
	BMS	6542.04		5680.28		10854.94		7774.80		1615.55		1165.15		3607.85		2408.74		4320.17		2342.94		9484.06		7576.26		3174.00		1914.53	
	WMS	97.47		117.28		127.96		135.13		124.38		143.62		113.68		110.85		100.03		133.21		111.02		125.84		98.64		107.55	
	F-ratio	67.12*		48.43*		84.83*		57.54*		12.99*		8.11*		31.74*		21.73*		43.19*		17.59*		85.43*		60.21*		32.18*		17.80*	
	F-S			0.90				0.84				0.82				0.85				0.85				0.91				0.94	
Grade 9: Tutees (2 sites)	Mean	73.00	73.00	76.72	1.96	73.00	75.97	73.61	-1.07													72.45	73.89	75.06	0.98				
Advanced, fall	SD	8.24		7.19		10.24		12.01														8.75		8.78					
Advanced, spring	N	20	(18)	18		20	(18)	18														20	(18)	18					
	BMS	26.32		14.67		1.05		61.34														20.84		52.71					
	WMS	70.20		54.06		110.61		149.31														79.67		78.64					
	F-ratio	0.37		0.27		0.01		0.41														0.76		0.67					
	F-S			0.91				0.66																0.84					

Key: Exp is the expected mean; SD = standard deviation; N = number of students (numbers in parentheses are the numbers of children for whom both fall and spring data are available); BMS = between mean square; WMS = within mean square; F-S is the fall to spring correlation.

*p = ≤ 0.05.

Table 3-14

DESCRIPTIVE STATISTICS FOR GRADE 1, BY PROJECT

Grade 1 Project	Statistic	Listening for Sounds Fall	Exp	Spring	t Test	Word Knowledge Fall	Exp	Spring	t Test	Reading Fall	Exp	Spring	t Test	Word Analysis Fall	Exp	Spring	t Test	Total Reading Fall	Exp	Spring	t Test	Total Math Fall	Exp	Spring	t Test
Catch-Up (4 sites)	Mean	25.52						30.60		24.18		34.57	-4.60			33.40				31.72		23.54		34.00	-11.05
Primer, fall	SD	3.65						8.55		4.32		6.19				6.87				6.86		5.09		5.45	
Primary I, spring	N	60						48		61	(47)	47				48				47		57	(46)	48	
	BMS	92.39						231.10		58.89		113.36				176.27				164.42		14.78		137.93	
	WMS	9.07						62.32		16.53		33.10				38.38				38.89		26.52		22.28	
	F ratio	10.19*						3.71*		3.56*		3.42*				4.59*				4.23*		0.56		6.19*	
	F-S											0.31												0.25	
Conquest (1 site)	Mean	32.64						41.90		30.93		42.05	-13.50			37.62				41.85					
Primer, fall	SD	4.56						7.56		4.56		4.30				4.15				4.32					
Primary I, spring	N	28	(21)					21		28	(20)	20				21				20					
	F-S											0.52													
PTR (2 sites)	Mean	23.56		29.04†				26.95‡		23.29		30.18	-18.77			29.24‡				28.50‡					
Primer, fall	SD	2.28		4.49				7.00		4.11		5.63				5.33				6.49					
Canton-Primer, spring	N	333		168				119		301	(242)	281				118				114					
Dallas-Primary I, spring	BMS	53.47								367.67		551.35													
	WMS	5.07								15.69		30.08													
	F ratio	10.55*								23.44*		18.33*													
	F-S											0.32													

Note: Test level was changed from Primer in fall to Primary I in spring for all grade 1 projects except Canton PTR, which received Primer in both fall and spring. Total Math represents the Numbers subtest score for Primer and the Math subtest for Primary I. Total Reading for spring includes Word Knowledge and Reading subtests. Norms used for Primary in spring are mid-year grade 1 norms.

Key: Exp is the expected mean; SD = standard deviation; N = number of students (numbers in parentheses are the number of children for whom both fall and spring data are available); BMS = between mean square; WMS = within mean square; F-S is the fall to spring correlation.

*p = ≤0.05.

†Canton only.

‡Dallas only.

Table 3-15

RESULTS OF THE NORM-REFERENCED ANALYSIS OF MAT STANDARD SCORES, BY GRADE, SITE, AND SUBTEST

a. Grades 2-5

Grade and Site	NR/NM	Word Knowledge: Gain over Fall	Word Knowledge: Gain over Expected Gain	Word Knowledge: t Test	Word Knowledge: Meets Criterion Growth	Word Knowledge: Normal Growth	Reading: Gain over Fall	Reading: Gain over Expected Gain	Reading: t Test	Reading: Meets Criterion Growth	Reading: Normal Growth	Total Reading: Gain over Fall	Total Reading: Gain over Expected Gain	Total Reading: t Test	Total Reading: Meets Criterion Growth	Total Reading: Normal Growth
Grade 2																
Catch-Up																
Bloomington	32/29	12.28	1.28	1.10	Unknown	Unknown	5.03	-2.13	-1.71	No	Unknown	8.97	-1.03	-1.08	No	Unknown
Brookport	5/5	16.60	5.20	2.05	Unknown	Unknown	4.60	-3.00	-0.96	Unknown	Unknown	11.20	2.20	1.03	Unknown	Unknown
Galax	10/9	9.60	-0.43	-0.28	No	Unknown	6.80	-1.69	-1.01	No	Unknown	8.50	-0.23	-0.23	No	Unknown
Wayne City	3/3*															
Conquest																
Benton Harbor	28/0*	15.89	4.89	4.83	Unknown	Yes	9.61	4.61	3.25	Unknown	Yes	13.57	4.50	4.67	Unknown	Yes
Cleveland	105/0*	9.01	-1.19	-1.92	No	Unknown	9.06	0.70	1.10	No	Unknown	9.25	0.39	0.93	No	Unknown
Gloversville	26/0*	8.96	-1.64	-1.83	No	Unknown	7.96	-0.04	-0.02	No	Unknown	8.38	-0.68	-0.73	No	Unknown
Grade 3																
Catch-Up																
Bloomington	22/30	2.76	-1.62	-1.74	No	Unknown	3.36	1.92	0.67	Unknown	Unknown	2.64	-0.01	-0.01	No	Unknown
Brookport	6/4*	5.00	2.08	1.38	Unknown	Unknown	5.83	3.93	1.09	Unknown	Unknown	4.33	1.50	0.62	Unknown	Unknown
Galax	8/6	1.88	-0.31	-0.08	Unknown	Unknown	4.63	3.25	0.67	Unknown	Unknown	1.38	-0.63	-0.15	Unknown	Unknown
Wayne City	4/4*															
Conquest																
Benton Harbor	24/0*	3.04	0.04	0.03	No	Unknown	4.92	2.92	2.23	Unknown	Yes	2.71	0.24	0.17	No	Unknown
Cleveland	94/0*	4.36	1.36	2.10	No	Yes	6.22	4.22	4.77	Unknown	Yes	4.42	2.35	3.33	No	Yes
Gloversville	7/0*	1.29	-1.71	-0.64	Unknown	Unknown	1.00	-0.00	-0.00	Unkown	Unknown	0.29	-2.51	-0.95	No	Unknown
IRIT																
Bloomington	13/0*	2.00	-2.50	-2.12	No	Unknown	7.54	5.87	2.90	Unknown	Yes	3.62	0.11	0.10	No	Unknown
Oklahoma City	28/0*	5.29	1.44	1.88	No	Unknown	5.93	4.72	3.80	Unknown	Yes	5.07	2.56	3.01	No	Yes
Schenectady	25/0*	6.70	3.20	2.60	Unknown	Yes	6.92	4.92	2.87	Unknown	Yes	6.00	3.35	2.47	Unknown	Yes
Grade 4																
Catch-Up																
Bloomington	40/39	3.00	-1.00	-1.09	No	Unknown	1.60	-3.34	-2.97	No	No	2.73	-2.28	-2.70	No	No
Brookport	8/6	3.25	-1.75	-0.57	Unknown	Unknown	3.50	-0.44	-0.26	No	Unknown	3.75	-1.25	-0.75	No	Unknown
Galax	7/7	3.71	-1.29	-0.45	Unknown	Unknown	2.29	-1.86	-0.30	Unknown	Unknown	4.43	0.43	0.11	Unknown	Unknown
Providence Forge	20/15	2.95	-1.43	-1.46	No	Unknown	2.80	-1.20	-1.06	No	Unknown	3.05	-1.95	-2.08	No	Unknown
Wayne City	8/8	4.00	-1.00	-0.95	No	Unknown	-2.25	-5.75	-4.36	No	No	1.88	-2.25	-2.64	No	No
Conquest																
Benton Harbor	28/0*	6.61	1.61	1.21	Unknown	Unknown	7.46	2.66	1.85	Unknown	Unknown	7.36	3.19	2.75	Unknown	Yes
Cleveland	53/0*	4.42	-0.58	-0.82	No	Unknown	3.59	-0.33	-0.44	No	Unknown	4.77	0.49	0.65	No	Unknown
Gloversville	27/0*	4.78	-0.22	-0.21	No	Unknown	4.37	-0.23	-0.17	No	Unknown	5.22	0.22	0.22	No	Unknown
IRIT																
Bloomington	38/0*	4.82	-0.18	-0.14	No	Unknown	3.75	-0.02	-0.01	No	Unknown	4.39	0.12	0.12	No	Unknown
Schenectady	6/0*	-1.00	-5.08	-2.78	No	No	-1.67	-6.50	-1.96	No	Unknown	-1.00	-6.00	-6.71	No	No
Grade 5																
Catch-Up																
Bloomington	32/26	4.16	0.12	0.10	No	Unknown	9.94	3.94	2.15	Unknown	Yes	6.94	1.94	1.56	Unknown	Unknown
Brookport	13/13	4.46	-0.54	-0.38	No	Unknown	11.23	5.23	3.01	Unknown	Yes	7.23	2.12	1.59	Unknown	Unknown
Galax	6/6	3.50	0.17	0.06	Unknown	Unknown	12.00	6.00	1.90	Unknown	Unknown	7.67	1.67	0.67	Unknown	Unknown
Providence Forge	18/16	3.94	-0.06	-0.04	No	Unknown	8.28	2.83	1.32	Unknown	Unknown	6.22	0.24	0.21	No	Unknown
Wayne City	4/4*															
Conquest																
Benton Harbor	20/0*	7.40	3.25	3.51	Unknown	Yes	8.35	3.27	1.82	Unknown	Unknown	8.60	3.60	2.35	Unknown	Yes
Cleveland	17/0*	8.06	4.05	2.98	Unknown	Yes	13.17	8.83	3.75	Unknown	Yes	10.12	5.12	2.80	Unknown	Yes
Gloversville	31/0*	5.06	1.06	1.45	No	Unknown	10.43	4.60	3.45	Unknown	Yes	7.97	2.65	3.69	No	Yes

Grade and Site	Math Computation: Gain over Fall	Math Computation: Gain over Expected Gain	Math Computation: t Test	Math Computation: Meets Criterion Growth	Math Computation: Normal Growth	Math Concepts: Gain over Fall	Math Concepts: Gain over Expected Gain	Math Concepts: t Test	Math Concepts: Meets Criterion Growth	Math Concepts: Normal Growth	Math Problem Solving: Gain over Fall	Math Problem Solving: Gain over Expected Gain	Math Problem Solving: t Test	Math Problem Solving: Meets Criterion Growth	Math Problem Solving: Normal Growth	Total Math: Gain over Fall	Total Math: Gain over Expected Gain	Total Math: t Test	Total Math: Meets Criterion Growth	Total Math: Normal Growth
Grade 2																				
Catch-Up																				
Bloomington																15.31	0.46	0.30	No	Unknown
Brookport																16.60	4.40	1.98	Unknown	Unknown
Galax																17.89	2.89	0.98	Unknown	Unknown
Wayne City																				
Conquest																				
Benton Harbor																				
Cleveland																				
Gloversville																				
Grade 3																				
Catch-Up																				
Bloomington	8.23	0.23	0.16	No	Unknown	2.97	-4.53	-3.09	No	No	4.60	-0.40	-0.19	No	Unknown	5.87	-2.13	-1.48	No	Unknown
Brookport																				
Galax	12.00	3.81	1.49	Unknown	Unknown	8.29	1.79	0.52	Unknown	Unknown	5.14	-0.86	-0.55	No	Unknown	8.67	2.67	1.41	Unknown	Unknown
Wayne City																				
Conquest																				
Benton Harbor																				
Cleveland																				
Gloversville																				
IRIT																				
Bloomington																				
Oklahoma City																				
Schenectady																				
Grade 4																				
Catch-Up																				
Bloomington	7.43	0.55	0.40	No	Unknown	6.13	0.46	0.31	No	Unknown	8.38	2.37	1.49	Unknown	Unknown	8.51	1.51	1.42	No	Unknown
Brookport	7.86	-0.14	-0.05	Unknown	Unknown	2.14	-3.86	-1.11	Unknown	Unknown	2.83	-2.17	-0.52	Unknown	Unknown	6.67	-0.83	-0.34	Unknown	Unknown
Galax	1.57	-5.57	-1.85	No	Unknown	4.71	-1.29	-0.59	Unknown	Unknown	3.00	-1.33	-0.30	Unknown	Unknown	3.57	-4.90	-1.54	No	Unknown
Providence Forge	9.32	2.55	1.34	Unknown	Unknown	7.58	1.58	1.14	Unknown	Unknown	3.69	-2.31	-1.02	No	Unknown	6.40	1.40	0.82	Unknown	Unknown
Wayne City	12.52	4.38	2.33	Unknown	Unknown	3.38	-2.63	-1.02	No	Unknown	1.13	-3.88	-2.44	No	No	6.25	-1.25	-1.03	No	Unknown
Conquest																				
Benton Harbor																				
Cleveland																				
Gloversville																				
IRIT																				
Bloomington																				
Schenectady																				
Grade 5																				
Catch-Up																				
Bloomington	10.57	6.07	3.11	Unknown	Yes	13.38	10.38	5.62	Yes	Yes	8.40	5.40	2.76	Unknown	Yes	12.73	9.52	4.84	Yes	Yes
Brookport	15.62	12.19	7.37	Yes	Yes	9.31	6.31	2.78	Unknown	Yes	10.34	7.36	3.10	Unknown	Yes	12.34	7.69	7.06	Yes	Yes
Galax	14.14	9.24	3.37	Unknown	Yes	5.43	2.43	0.55	Unknown	Unknown	9.83	6.83	1.66	Unknown	Unknown	12.17	8.54	2.05	Unknown	Unknown
Providence Forge	10.94	4.94	3.00	Unknown	Yes	6.53	3.78	2.42	Unknown	Yes	5.56	2.56	1.21	Unknown	Unknown	8.19	3.73	3.45	Unknown	Yes
Wayne City																				
Conquest																				
Benton Harbor																				
Cleveland																				
Gloversville																				

Table 3-15 (Continued)

b. Grades 6-9

Grade and Site	NR/NM	Word Knowledge					Reading					Total Reading				
		Gain over Fall	Gain over Expected Gain	t Test	Meets Criterion Growth	Normal Growth	Gain over Fall	Gain over Expected Gain	t Test	Meets Criterion Growth	Normal Growth	Gain over Fall	Gain over Expected Gain	t Test	Meets Criterion Growth	Normal Growth
Grade 6																
Catch-Up																
Brookport	8/8	2.00	-1.00	-0.68	No	Unknown	10.12	5.00	1.58	Unknown	Unknown	5.75	1.75	3.33	No	Yes
Galax	8/9	4.38	1.37	1.27	No	Unknown	9.63	4.62	4.40	Unknown	Yes	7.25	3.19	3.15	Unknown	Yes
Providence Forge	18/16	1.44	-1.56	-1.01	No	Unknown	2.95	-2.47	-1.35	No	Unknown	2.11	-2.03	-1.28	No	Unknown
Wayne City	5/5	1.60	-2.40	-1.37	No	Unknown	4.20	-0.80	-0.19	Unknown	Unknown	2.60	-1.20	-0.52	Unknown	Unknown
Conquest																
Benton Harbor	14/0*	4.64	1.39	0.87	Unknown	Unknown	6.07	0.71	0.41	Unknown	Unknown	5.86	1.86	1.17	Unknown	Unknown
Cleveland	22/0*	7.09	4.09	3.29	Unknown	Yes	10.64	4.91	2.60	Unknown	Yes	8.95	5.94	5.75	Unknown	Yes
Gloversville	51/0*	4.82	9.98	3.15	No	Unknown	6.65	3.65	1.39	No	Unknown	6.39	2.39	2.91	No	Yes
HIT																
Lexington																
Tutors	22/10	5.73	1.73	1.69	No	Unknown	3.68	-1.32	-0.99	No	Unknown	5.59	1.59	1.86	No	Unknown
Tutees	41/23	4.12	0.05	0.05	No	Unknown	7.12	0.19	0.14	No	Unknown	5.71	1.29	1.38	No	Unknown
Grade 7																
Catch-Up																
Wayne City	5/5	1.20	0.40	0.17	Unknown	Unknown	-0.20	-0.60	-0.30	No	Unknown	0.00	0,00	0,00	No	Unknown
HIT																
Lexington																
Tutors	24/9	1.88	0.88	0.87	No	Unknown	-1.67	-2.67	-2.90	No	No	-0.08	-0.08	-0.09	No	Unknown
Tutees	20/8	-0.95	-0.95	-0.43	No	Unknown	1.70	-0.30	-0.11	No	Unknown	-0.25	-1.25	-0.64	No	Unknown
Olean																
Tutors	66/63	2.03	-0.72	-0.90	No	Unknown	2.62	1.62	1.66	No	Unknown	2.59	0.59	0.88	No	Unknown
Tutees	44/40	1.39	0.54	0.56	No	Unknown	-0.41	-1.90	-1.68	No	Unknown	0.59	0.59	0.70	No	Unknown
Grade 8																
Catch-Up																
Wayne City	4/4*															
HIT																
Lexington																
Tutors	31/18	3.00	1.82	1.74	No	Unknown	2.68	0,68	0.46	No	Unknown	3.19	2.19	2.12	No	Yes
Tutees	17/0*	-0.12	-0.41	-0.26	No	Unknown	3.33	1.89	0.96	Unknown	Unknown	1.71	0,43	0.32	No	Unknown
Olean																
Tutors	15/1*	4.27	2.74	1.69	Unknown	Unknown	4.13	3.46	1.94	Unknown	Unknown	4.87	3.87	3.29	Unknown	Yes
Tutees	45/40	3.93	2.45	2.37	No	Yes	2.87	1.76	1.32	No	Unknown	3.82	2.82	2.90	No	Yes
R-3																
Charlotte	228/214	3.81	1.81	3.73	No	Yes	4.71	4.71	8.72	Unknown	Yes	4.48	4.48	10.82	No	Yes
Lake Village	123/120	3.29	1.51	2.20	No	Yes	1.92	0.99	1.19	No	Unknown	2.98	1.98	3.22	No	Yes
Lorain	271/285	3.80	1.90	4.43	No	Yes	4.94	4.94	9.14	Unknown	Yes	4.77	4.32	11.28	No	Yes
Schenectady	157/155	3.90	1.90	3.97	No	Yes	4.35	4.35	6.60	No	Yes	4.64	3.72	6.48	No	Yes
Grade 9																
HIT																
Lexington																
Tutors	40/17	0.53	-1.48	-1.33	No	Unknown	3.63	0.14	0.10	No	Unknown	2.07	1.07	0.98	No	Unknown
Tutees	17/0*	3.65	1.65	1.78	No	Unknown	2.12	-1.91	-0.84	No	Unknown	3.35	1.35	1.08	No	Unknown
Olean																
Tutors	31/23	0.90	-1.32	-1.69	No	Unknown	2.41	-0.52	-0.34	No	Unknown	1.61	-0.35	-0.36	No	Unknown
Tutees	1/0*															

Grade and Site	Math Computation					Math Concepts					Math Problem Solving					Total Math				
	Gain over Fall	Gain over Expected Gain	t Test	Meets Criterion Growth	Normal Growth	Gain over Fall	Gain over Expected Gain	t Test	Meets Criterion Growth	Normal Growth	Gain over Fall	Gain over Expected Gain	t Test	Meets Criterion Growth	Normal Growth	Gain over Fall	Gain over Expected Gain	t Test	Meets Criterion Growth	Normal Growth
Grade 6																				
Catch-Up																				
Brookport	15.75	13.37	3.90	Yes	Yes	9.25	6.44	3.56	Unknown	Yes	11.62	8.62	2.51	Unknown	Yes	13.75	9.75	4.20	Yes	Yes
Galax	16.56	13.22	9.76	Yes	Yes	12.31	9.11	2.67	Unknown	Yes	5.00	1.28	0.81	Unknown	Unknown	12.11	9.06	4.97	Yes	Yes
Providence Forge	7.72	2.72	1.56	Unknown	Unknown	5.88	2.88	1.52	Unknown	Unknown	7.94	3.04	1.91	Unknown	Unknown	8.63	5.00	4.40	Unknown	Yes
Wayne City	11.60	7.60	1.61	Unknown	Unknown	9.60	6.60	6.13	Unknown	Yes	3.40	1.50	0.65	Unknown	Unknown	8.80	5.00	1.92	Unknown	Unknown
Conquest																				
Benton Harbor																				
Cleveland																				
Gloversville																				
HIT																				
Lexington																				
Tutors	8.36	3.90	1.70	Unknown	Unknown	5.09	2.09	1.29	Unknown	Unknown	6.10	3.10	0.90	Unknown	Unknown	6.90	2.90	1.46	Unknown	Unknown
Tutees	12.92	10.92	7.52	Yes	Yes	9.35	10.45	7.68	Yes	Yes	7.40	4.60	4.14	Unknown	Yes	11.40	7.04	6.05	Yes	Yes
Grade 7																				
Catch-Up																				
Wayne City	5.20	2.30	0.89	Unknown	Unknown	-4.20	-4.70	-0.95	Unknown	Unknown	10.60	8.60	1.28	Unknown	Unknown	4.60	3.10	0.93	Unknown	Unknown
HIT																				
Lexington																				
Tutors	9.56	8.67	3.47	Unknown	Yes	0.44	0.33	0.14	Unknown	Unknown	10.50	7.70	4.21	Unknown	Yes	7.33	5.33	4.44	Unknown	Yes
Tutees	9.88	9.31	4.19	Unknown	Yes	1.88	-2.13	-0.86	No	Unknown	3.50	-0.35	-0.15	Unknown	Unknown	5.63	4.17	2.87	Unknown	Yes
Olean																				
Tutors	3.74	1.09	1.33	No	Unknown	0.75	-1.25	-1.24	No	Unknown	8.49	5.49	6.87	Unknown	Yes	5.44	3.44	5.22	Unknown	Yes
Tutees	4.55	2.55	2.07	Unknown	Yes	1.57	1.57	1.23	No	Unknown	7.60	5.60	3.49	Unknown	Yes	5.08	4.07	3.97	Unknown	Yes
Grade 8																				
Catch-Up																				
Wayne City																				
HIT																				
Lexington																				
Tutors	-0.06	0.94	0.35	Unknown	Unknown	-3.00	-3.00	-1.80	No	Unknown	2.39	0.89	0.42	Unknown	Unknown	-0.83	-0.83	-0.40	No	Unknown
Tutees																				
Olean																				
Tutors																				
Tutees	3.69	4.69	3.89	Unknown	Yes	1.91	1.91	1.30	No	Unknown	2.84	1.84	1.86	No	Unknown	2.75	2.75	2.56	Unknown	Yes
R-3																				
Charlotte	4.97	5.97	11.33	Yes	Yes	3.25	3.25	5.88	No	Yes	4.06	3.14	5.77	No	Yes	4.18	4.18	10.33	Unknown	Yes
Lake Village	1.33	2.33	3.40	No	Yes	1.83	1.83	2.39	No	Yes	1.23	0.95	1.17	No	Unknown	1.49	1.49	2.84	No	Yes
Lorain	2.67	2.78	7.21	No	Yes	3.03	3.03	7.09	No	Yes	4.03	3.03	6.11	No	Yes	3.41	3.41	10.24	No	Yes
Schenectady	1.53	1.07	1.88	No	Unknown	5.23	5.23	8.81	Unknown	Yes	3.54	2.97	4.01	No	Yes	3.69	3.69	7.43	No	Yes
Grade 9																				
HIT																				
Lexington																				
Tutors	3.84	3.64	2.40	Unknown	Yes	4.71	4.71	3.39	Unknown	Yes	6.21	5.05	2.75	Unknown	Yes	5.18	5.18	6.39	Unknown	Yes
Tutees																				
Olean																				
Tutors	4.69	1.90	1.31	No	Unknown	2.48	2.98	1.94	Unknown	Unknown	3.04	1.47	1.13	No	Unknown	3.25	3.26	3.35	Unknown	Yes
Tutees																				

Table 3-15 (Concluded)

c. Grade 1

Grade and Site	N	Listening for Sounds/Words Analysis: Gain Over Fall	Gain Over Expected Gain	t Test	Meets Criterion Growth	Meets Normal Growth	N	Reading: Gain Over Fall	Gain over Expected Gain	t Test	Meets Criterion Growth	Meets Normal Growth	N	Total Math: Gain Over Fall	Gain over Expected Gain	t Test	Meets Criterion Growth	Meets Normal Growth
Grade 1																		
Catch-Up																		
Bloomington	26	7.96	-3.75	-3.04	No	No	26	10.58	-3.76	-3.14	No	No	25	8.44	-12.36	-9.94	No	No
Brookport	6	4.00	-9.86	-8.76	No	No	5	8.00	-6.50	-3.82	No	No	6	13.33	-8.67	-7.21	No	No
Galax	8	6.00	-5.29	-4.73	No	No	8	8.13	-6.38	-3.25	No	No	7	16.57	-3.43	-1.03	Unknown	Unknown
Wayne City	7	9.29	-5.95	-2.89	No	No	8	12.87	-4.43	-1.47	No	Unknown	8	8.13	-13.38	-15.67	No	No
Conquest																		
Cleveland	21	4.86	-12.90	-15.06	No	No	20	10.80	-13.70	-13.50	No	No						
PTR																		
Canton	165	5.85	-0.32	-0.97	No	Unknown	131	7.28	-0.18	-0.50	No	Unknown						
Dallas	112	5.45	-5.90	-13.12	No	No	111	7.41	-7.09	-10.22	No	No						

NOTES: NR = number of children for whom Total Reading scores are available for both fall and spring; NM = number of children for whom Total Math scores are available for both fall and spring; N = number of children for whom test scores are available for both fall and spring.

Test level was changed from Primer in fall (Listening for Sounds subtest) to Primary I in spring (Word Analysis subtest) for all grade 1 projects except Canton PTR, which received the Primer in both fall and spring. Gains for Total Math are based on the Numbers subtest score for Primer and Math subtest for Primary I. Norms used for Primer in spring are mid-year grade 1 norms. Norms used for Primary I in spring are end-of-year grade 1 norms.

*Data are not shown for grades having fewer than five students with valid test data.

It is noteworthy that these gains were made by students whose averages were well below the norm group's average fall performance. Tables 3-16 through 3-21 show the means and corresponding interpolated percentiles for each project with more than four cases per grade, by grade within PIP. In general we see that, as intended, the PIP projects served very low achievers. In Olean HIT, most children who were performing well above their grade's norm group average were tutors for the project. At most projects, most grades showed averages less than the 25th percentile of the norm group, with several grades showing averages as low as the 2nd percentile. Exceptions were the two PTR projects, which showed percentiles of 38 and 40 in Canton and in Dallas, respectively. For students with valid tests, this is a deviation from package specifications in that PTR was for children in the bottom quartile. Of course, if we were to take into account the unmeasured performance of those children who could not respond to the MAT, the "true percentile" would be lower.

A striking feature of the data in Tables 3-16 through 3-21 is the evidence for "grade effects" in the Reading subtest. Gains and losses in mean percentiles for Reading are shown in Table 3-22 by site and grade. Three PIPs with seven sites had first grades, all of which showed fall to spring losses in the percentile of their means. In the fourth grade of the three PIPs with ten sites, nine showed losses; one site at this grade, Benton Harbor Conquest, showed a gain. In the third grade of the three PIPs with nine sites, eight showed gains in the percentile of the mean; Gloversville Conquest at this grade was stationary. In the fifth grade two PIPs with seven sites all showed gains. At the sixth grade, six of nine groups in three PIPs showed gains in the percentile of their means. At the eighth grade all sites showed gains.

Again, there may be a problem with the norming and linking of the members of the MAT battery. At the first and fourth grades, we see definite decreases across a variety of sites and instructional programs. At the third, fifth, and eighth grades we see increases across a variety of sites and instructional programs.

Insofar as these analyses are concerned, we can only hope that these effects represent some feature associated with the PIPs. However, of the six sites with a third, fourth, and fifth grade, all but Benton Harbor showed gains in Reading percentile at the third and fifth grades and losses at the fourth. Gloversville remained stationary at the third grade. All five sites with a third, fourth, fifth, and sixth grade confirmed the above sequence and showed sixth grade gains.

The possibility of grade effects further confirms our case for discounting the norm-referenced analysis. However, within the context of the norm-referenced analysis, we must accept the effects not as artifacts, but as evidence, for example, that there was something wrong in five out of six sites' fourth grades, which was not wrong in these same sites' third grades.

At this time we know of no such grade-related problems, and we infer that the grade effects are artifacts caused by some defect in the MAT norms. The effects are discussed further in Section 6.3.

We conclude on the basis of available data that by and large the PIP projects did not pass the norm-referenced criterion of educationally significant growth.

3.7.2 Comparison with Dissemination and Review Panel Criteria

The generally negative results presented above do not show that the PIP field-test projects have failed the criterion that the originating projects passed. As discussed in Section 2, the criterion used at the original sites, of necessity, changed from test to test and from grade to grade. That is, the "one-third standard deviation" does not represent a single criterion; there were multiple criteria.

However, while with few exceptions the PIP projects have failed the general criterion set at the beginning of this evaluation, it has not been demonstrated that they failed the criterion of replicating the exemplary programs' effectiveness, since the designs for evaluating the exemplary programs and the PIPs were not identical.

In the next paragraphs, we sketch what the criterion analysis for each of the PIPs would have looked like, had the procedures for the original sites been followed.

- Catch-Up--The original program was evaluated on 1971-72 data using the Cooperative Primary Reading Test for grades 1, 2, and 3. The project passed the one-third standard deviation criterion on this test at grades 1 and 2. In math it passed at grades 1 and 3.

  For 1971-72 data on grades 4, 5, and 6, the California Test of Basic Skills (CTBS) was used for evaluation purposes. Only grade 4 passed either reading or math. Grades 5 and 6 did not pass either.

  In 1972-73, the MAT was used for grades 1, 2, and 3. In an evaluation based on these data, grade 3 passed reading. The other grades provided inadequate data for reaching a decision. For this year, the CTBS was again used for

Table 3-16

MEANS AND INTERPOLATED PERCENTILES OF MEANS FOR CATCH-UP
FOR FALL AND SPRING, BY PROJECT AND SUBTEST

Catch-Up	NR/NM	Word Knowledge				Reading				Total Reading				Math Computation				Math Concepts				Math Problem Solving				Total Math			
		Fall		Spring		Fall		Spring		Fall		Spring		Fall		Spring		Fall		Spring		Fall		Spring		Fall		Spring	
		Mean	Percentile	Mean	Percentile	Mean	Percentile	Mean	Percentile	Mean	Percentile	Mean	Percentile	Mean	Percentile	Mean	Percentile	Mean	Percentile	Mean	Percentile	Mean	Percentile	Mean	Percentile	Mean	Percentile	Mean	Percentile
Bloomington																													
Grade 2	32/29																												
Primary I		34.62	13			35.16	18			34.44	15															32.86	12		
Primary II				46.91	16			40.19	14			43.41	13															48.[illegible]7	12
Grade 3	22/30																												
Primary II		51.39	26			45.86	12			47.73	17			48.39	17			53.47	34			49.86	18			52.97	20		
Elementary				54.15	22			49.23	16			50.36	17			56.61	17			56.43	21			54.46	17			58.83	16
Grade 4	40/39																												
Elementary		55.80	18	58.80	16	53.13	18	54.72	12	53.15	18	55.87	14	56.20	11	63.62	13	54.67	13	60.80	14	54.00	10	62.37	13	57.33	9	65.85	11
Grade 5	32/26																												
Elementary		61.94	16			56.22	11			58.22	12			59.50	3			54.10	2			58.23	5			59.62	1		
Intermediate				66.04	16			66.16	17			65.16	16			70.07	10			67.48	13			66.63	[illegible]			72.35	9
Brookport																													
Grade 2	5/5																												
Primary I		30.20	7			35.60	19			32.00	10															30.60	7		
Primary II				46.80	16			40.20	14			43.20	12															47.20	11
Grade 3	6/4*																												
Primary II		44.17	8			44.67	11			43.83	10																		
Elementary				49.17	11			50.50	19			48.17	12																
Grade 4	8/6																												
Elementary		57.25	23	60.50	19	57.12	26	60.63	25	55.75	24	59.50	21	62.86	29	70.71	29	57.14	16	59.29	11	59.67	17	62.50	13	63.67	21	70.33	19
Grade 5	13/13																												
Elementary		59.31	12			52.77	9			55.75	10			64.15	7			56.15	3			57.29	5			62.85	4		
Intermediate				63.77	11			64.00	14			63.00	12			79.77	33			65.46	9			67.6[illegible]	13			76.38	17
Grade 6	8/8																												
Elementary		66.50	15			63.13	12			64.38	13			65.75	3			61.62	4			60.25	3			66.[illegible]	2		
Intermediate				68.50	13			73.25	21			70.12	15			81.50	19			70.87	10			71.88	10			80.00	12
Galax																													
Grade 2	10/9																												
Primary I		40.90	32			37.50	23			38.80	26															35.89	20		
Primary II				50.50	29			44.30	21			47.30	24															53.78	33
Grade 3	8/6																												
Primary II		45.63	10			40.75	7			42.87	8			42.43	8			43.00	9			44.14	9			46.17	7		
Elementary				47.50	9			45.37	9			44.25	7			54.43	13			51.2[illegible]	11			49.29	8			54.83	10
Grade 4	7/7																												
Elementary		49.86	9	53.57	7	48.71	10	51.00	7	46.71	7	51.14	7	51.29	6	52.86	2	51.57	9	56.29	7	50.00	5	53.00	4	52.57	5	56.14	2
Grade 5	6/6																												
Elementary		56.33	7			54.50	10			54.00	8			57.86	2			56.86	4			59.00	6			60.17	2		
Intermediate				59.63	8			66.50	18			61.67	11			72.00	12			62.29	6			68.83	16			72.33	9
Grade 6	8/9																												
Elementary		64.00	10			61.75	11			61.88	10			67.11	4			64.00	6			72.56	17			71.89	6		
Intermediate				68.37	13			71.37	17			69.12	14			83.67	23			76.11	20			77.56	19			84.00	20
Providence Forge																													
Grade 4	20/15																												
Elementary		54.25	15	57.20	12	52.00	14	54.80	12	51.70	15	54.75	12	56.37	11	65.68	18	53.00	10	60.58	13	55.87	12	59.56	10	57.00	9	65.40	10
Grade 5	18/16																												
Elementary		63.44	19			62.11	19			52.06	20			68.06	16			65.53	15			67.62	19			71.12	14		
Intermediate				67.39	19			70.39	25			68.28	21			79.00	30			72.06	26			73.19	24			79.31	23
Grade 6	18/16																												
Elementary		64.28	11			60.37	10			61.72	10			70.33	8			64.76	7			65.31	8			70.75	5		
Intermediate				65.72	8			63.32	6			63.83	7			78.06	11			70.65	10			73.25	13			79.38	11
Wayne City																													
Grade 2	3*/3*																												
Primary I																													
Primary II																													
Grade 3	4*/4*																												
Primary II																													
Elementary																													
Grade 4	8/8																												
Elementary		60.12	30	64.12	27	58.00	27	55.75	14	5[illegible].75	28	59.62	21	63.25	31	75.87	44	57.25	17	60.63	13	60.25	19	61.37	11	63.75	21	70.00	[illegible]
Grade 5	4*/4*																												
Elementary																													
Intermediate																													
Grade 6	5/5																												
Elementary		72.40	29			74.60	32			72.40	29			81.00	28			74.60	23			81.20	34			82.40	25		
Intermediate				74.00	22			78.80	30			76.00	26			92.60	46			84.20	38			84.60	39			91.20	43
Grade 7	5/5																												
Intermediate		78.40	29			82.20	36			80.40	29			91.20	38			93.40	70			86.60	31			94.20	4[illegible]		
Advanced				79.60	30			82.00	34			80.40	29			96.40	47			89.20	50			97.20	52			98.80	51
Grade 8	4*/4*																												
Advanced																													

NOTE: NR = Number of children for whom Total Reading scores are available for both fall and spring; NM = number of children for whom Total Math scores are available for both fall and spring. Other subtest means and percentiles may be based on larger sample sizes.

*Data are not shown for grades having fewer than five students with valid test data.

Table 3-17

MEANS AND INTERPOLATED PERCENTILES OF MEANS FOR CONQUEST
FOR FALL AND SPRING, BY PROJECT AND SUBTEST

Conquest	Total N	Word Knowledge Fall Mean	Word Knowledge Fall Percentile	Word Knowledge Spring Mean	Word Knowledge Spring Percentile	Reading Fall Mean	Reading Fall Percentile	Reading Spring Mean	Reading Spring Percentile	Total Reading Fall Mean	Total Reading Fall Percentile	Total Reading Spring Mean	Total Reading Spring Percentile
Benton Harbor													
Grade 2	28												
Primary I		33.18	11			32.61	[illegible]			32.07	10		
Primary II				49.07	22			42.21	18			45.64	19
Grade 3	24												
Primary II		48.21	14			43.20	9			45.08	11		
Elementary				51.25	15			48.12	14			47.79	12
Grade 4	28												
Elementary		51.07	10	57.68	13	50.39	12	57.86	18	49.18	10	56.54	15
Grade 5	20												
Elementary		58.15	10			59.85	16			57.40	11		
Intermediate				65.55	15			68.20	20			66.00	17
Grade 6	14												
Elementary		63.50	10			63.36	13			62.21	10		
Intermediate				68.14	12			69.43	13			68.07	12
Cleveland													
Grade 2	105												
Primary I		40.43	30			37.84	24			38.59	25		
Primary II				49.44	24			46.94	26			47.84	27
Grade 3	64												
Primary II		48.48	15			44.42	10			45.75	12		
Elementary				52.87	19			50.63	19			50.23	17
Grade 4	53												
Elementary		52.04	12	56.45	11	49.12	10	52.71	9	49.30	11	54.08	11
Grade 5	17												
Elementary		52.18	3			49.17	4			49.71	3		
Intermediate				60.24	8			62.33	11			59.82	8
Grade 6	22												
Elementary		56.41	3			53.45	3			53.95	2		
Intermediate				63.50	6			64.09	7			62.91	6
Gloversville													
Grade 2	26												
Primary I		38.81	24			36.46	21			37.19	21		
Primary II				47.77	19			44.42	21			45.56	18
Grade 3	[illegible]												
Primary I		49.14	17			47.43	15			47.43			
Elementary				50.43	13			48.43	15			47.71	11
Grade 4	27												
Elementary		[illegible]	12	56.52	11	52.63	16	57.00	16	50.26	13	55.46	13
Grade 5	31												
Elementary		[illegible]	19			58.29	14			60.32	17		
Intermediate				68.65	23			68.74	22			68.29	21
Grade 6	51												
Elementary		70.31	23			70.04	24			69.59	21		
Intermediate				75.14	26			76.69	27			75.98	26

Note: Total N = number of children for whom Total Reading scores are available for both fall and spring. Other subtest means and percentiles may be based on larger sample sizes.

grades 4, 5, and 6. Grade 5 passed reading and grades 4 and 5 passed math. Thus, at one time or another, all grades except grade 1 failed to meet the criterion.

- Conquest--Conquest was not examined on the basis of a normed test.
- HIT--Grade 6 was evaluated for HIT on the basis of 1971-72 Wide Range Achievement Test data. The HIT pupils were judged against the mean change of the age-specific norms.
- IRIT--IRIT was evaluated using 1972-73 data from the California Achievement Test. Grade 3 was successful. Data were inadequate for evaluation of fourth graders.
- PTR--For five sites, including the originating site, previous PTR evaluation data are available for tutored and control groups. The data show statistically significant results in favor of the tutored groups.
- R-3--R-3 was evaluated using 1970-71 and 1971-72 data on the CTBS. Eighth grade children's scores were examined and the two-year gains in reading met the criterion. Gains in math did not.

  Data collected in 1972-73 on seventh graders indicated that reading gains did not meet the criterion, but math gains did.

Thus, in only one project, Catch-Up, was the MAT battery used and then only for grades 1 through 3 during 1972-73. Only grade 3 met the criterion.

Whether the PIP projects could meet the original criteria is an open question. Based on the analysis (in Section 5) of the similarity of the original validating tests to the MAT, it seems possible that, if the original programs had been tested with the MAT and had been tried on the PIP criteria, they might not have fared much better than the PIP field-test projects.

3.7.3 Conclusion

The claim that the six PIPs would induce projects that could pass the norm-referenced analysis for the achievement of educationally significant growth is one of the PIPs' chief features. In Section 2, we reviewed with the reader the foundations for this claim; we also discussed the consequences of uncritically accepting the one-third standard deviation criterion or the equipercentile definition of expected growth.

Table 3-18

MEANS AND INTERPOLATED PERCENTILES OF MEANS FOR HIT
FOR FALL AND SPRING, BY PROJECT AND SUBTEST

		Word Knowledge				Reading				Total Reading			
		Fall		Spring		Fall		Spring		Fall		Spring	
HIT	NR/NM	Mean	Percentile	Mean	Percentile	Mean	Percentile	Mean	Percentile	Mean	Percentile	Mean	Percentile
Lexington													
Grade 6													
Elementary													
Tutors	22/10	69.55	21			72.27	27			69.95	22		
Tutees	41/23	58.07	5			58.10	8			56.83	5		
Intermediate													
Tutors	22/10			75.27	27			75.95	26			75.55	25
Tutees	41/23			62.20	5			65.22	8			62.54	6
Grade 7													
Intermediate													
Tutors	24/9	77.79	28			80.29	33			79.25	28		
Tutees	20/8	65.45	7			66.90	11			65.35	9		
Advanced													
Tutors	24/9			79.67	31			78.62	27			79.17	28
Tutees	20/8			64.50	7			68.60	11			65.10	7
Grade 8													
Advanced													
Tutors	31/18	75.65	16	78.65	18	73.71	15	76.39	16	74.35	13	77.55	16
Tutees	17/0*	67.41	6	67.29	5	65.17	5	68.50	7	65.18	5	66.88	6
Grade 9													
Advanced													
Tutors	40/17	80.10	16	80.63	15	74.02	12	77.65	12	77.37	12	79.45	13
Tutees	17/0*	73.29	9	76.94	11	71.94	10	74.06	7	72.12	8	75.47	9
Olean													
Grade 7													
Intermediate													
Tutors	66/63	89.74	65			88.36	53			89.74	60		
Tutees	44/40	76.70	25			76.02	25			76.45	24		
Advanced													
Tutors	66/63			91.77	64			90.98	58			92.33	61
Tutees	44/40			78.09	26			75.61	23			77.05	24
Grade 8													
Advanced													
Tutors	15/1*	96.93	62	101.20	67	94.67	55	98.80	63	96.80	60	101.67	67
Tutees	45/40	80.50	23	84.43	27	76.78	18	79.64	21	78.82	18	82.64	23
Grade 9													
Advanced													
Tutors	31/23	97.23	55	98.13	51	96.16	50	98.56	50	98.06	54	99.68	53
Tutees	1/0*												

	Math Computation				Math Concepts				Math Problem Solving				Total Math			
	Fall		Spring		Fall		Spring		Fall		Spring		Fall		Spring	
HIT	Mean	Percentile	Mean	Percentile	Mean	Percentile	Mean	Percentile	Mean	Percentile	Mean	Percentile	Mean	Percentile	Mean	Percentile
Lexington																
Grade 6																
Elementary																
Tutors	80.09	26			70.27	15			79.00	30			80.80	22		
Tutees	60.08	1			55.35	1			59.04	2			61.65	1		
Intermediate																
Tutors			88.45	37			75.36	19			85.10	40			87.70	31
Tutees			73.00	5			64.69	4			66.64	5			72.70	3
Grade 7																
Intermediate																
Tutors	80.89	16			77.78	20			76.40	12			83.44	15		
Tutees	71.75	5			66.12	5			67.30	4			71.62	2		
Advanced																
Tutors			90.44	31			78.22	20			86.90	28			90.78	28
Tutees			81.62	16			68.00	3			70.80	4			77.25	6
Grade 8																
Advanced																
Tutors	90.28	21	90.22	22	80.39	15	77.39	11	84.00	17	86.39	19	90.17	16	89.33	15
Tutees																
Grade 9																
Advanced																
Tutors	99.21	31	103.05	40	82.24	13	86.95	20	90.68	18	96.89	24	96.00	22	101.18	32
Tutees																
Olean																
Grade 7																
Intermediate																
Tutors	101.72	79			96.88	78			93.83	52			102.03	68		
Tutees	86.76	28			76.81	22			80.00	18			86.30	19		
Advanced																
Tutors			105.46	83			97.63	75			102.32	67			107.48	77
Tutees			91.31	33			80.38	25			87.60	29			91.37	29
Grade 8																
Advanced																
Tutors																
Tutees	89.73	20	93.42	25	81.41	17	83.33	19	89.05	24	91.89	30	91.05	18	93.80	24
Grade 9																
Advanced																
Tutors	106.42	57	111.12	62	99.00	50	101.48	57	106.76	58	109.80	62	109.96	54	113.22	62
Tutees																

Note: NR = number of children for whom Total Reading scores are available for both fall and spring; NM = number of children for whom Total Math scores are available for both fall and spring. Other subtest means and percentiles may be based on larger sample sizes.

*Data are not shown for grades having fewer than five students with valid test data.

Table 3-19

MEANS AND INTERPOLATED PERCENTILES OF MEANS FOR IRIT
FOR FALL AND SPRING, BY PROJECT AND SUBTEST

IRIT	Total N	Word Knowledge				Reading				Total Reading			
		Fall		Spring		Fall		Spring		Fall		Spring	
		Mean	Percentile	Mean	Percentile	Mean	Percentile	Mean	Percentile	Mean	Percentile	Mean	Percentile
Bloomington													
Grade 3	13												
Primary II		55.14	39			49.69	21			52.15	33		
Elementary				57.14	30			57.23	39			55.77	33
Grade 4	28												
Elementary		59.50	28	64.32	27	57.46	27	61.21	26	57.46	27	61.86	27
Oklahoma City													
Grade 3	28												
Primary II		50.86	23			46.41	13			48.04	18		
Elementary				56.14	27			52.34	23			53.11	24
Schenectady													
Grade 3	25												
Primary II		47.56	13			42.68	9			44.72	11		
Elementary				53.76	23			49.60	17			50.72	19
Grade 4	6												
Elementary		54.83	16	53.83	8	53.33	18	51.67	8	53.00	18	52.00	8

Note: Total N = number of children for whom Total Reading scores are for available for both fall and spring. Other subtest means and percentiles may be based on larger sample sizes.

Table 3-20

MEANS AND INTERPOLATED PERCENTILES OF MEANS FOR GRADE ONE
FOR FALL AND SPRING, BY PROJECT AND SUBTEST

Project	Total N	Listening for Sounds		Word* Analysis		Total N	Reading				Total N	Numbers		Mathematics	
		Fall		Spring			Fall		Spring			Fall		Spring	
		Mean	Percentile	Mean	Percentile		Mean	Percentile	Mean	Percentile		Mean	Percentile	Mean	Percentile
Catch-Up															
Bloomington	26					26					25				
Primer		24.58	42				23.23	41				23.80	55		
Primary I				32.54	21				33.81	29				32.24	17
Brookport	6					5					6				
Primer		28.33	73				25.00	50				25.17	63		
Primary I				32.33	20				33.00	24				38.50	41
Galax	8					8					7				
Primer		23.87	35				24.12	47				23.00	52		
Primary I				29.87	14				32.25	23				39.57	44
Wayne City	7					8					8				
Primer		30.57	86				27.50	73				24.25	58		
Primary I				39.86	63				40.37	54				32.37	17
Conquest															
Cleveland	21					20									
Primer		32.76	92				31.25	93							
Primary I				37.62	51				43.05	62					
PTR															
Canton	165					131									
Primer		23.18	29				22.18	38							
Primer				29.03	28†				29.47	37†					
Dallas	112					111									
Primer		24.01	36				24.62	49							
Primary I				29.46	17				32.04	22					

Note: Total N = number of children for whom subtest scores are available for both fall and spring. No Total Reading score.

*Listening for Sounds for Canton only.

† Mid-year (not end-of-year) percentiles.

Table 3-21

MEANS AND INTERPOLATED PERCENTILES OF MEANS FOR R-3
FOR FALL AND SPRING, BY PROJECT AND SUBTEST

a. Reading

R-3	NR	Word Knowledge				Reading				Total Reading			
		Fall		Spring		Fall		Spring		Fall		Spring	
		Mean	Percentile	Mean	Percentile	Mean	Percentile	Mean	Percentile	Mean	Percentile	Mean	Percentile
Charlotte Grade 8, Advanced	228	83.61	29	87.42	35	63.81	27	88.52	39	84.12	26	88.60	36
Lake Village Grade 8, Advanced	123	74.43	14	77.72	17	77.15	18	79.07	20	75.22	14	78.20	16
Lorain Grade 8, Advanced	271	82.41	27	86.21	31	81.21	23	86.15	32	82.09	23	86.87	32
Schenectady Grade 8, Advanced	157	81.51	25	85.41	29	80.54	23	84.89	28	81.15	22	85.79	30

b. Math

R-3	NM	Math Computation				Math Concepts				Math Problem Solving				Total Math			
		Fall		Spring		Fall		Spring		Fall		Spring		Fall		Spring	
		Mean	Percentile	Mean	Percentile	Mean	Percentile	Mean	Percentile	Mean	Percentile	Mean	Percentile	Mean	Percentile	Mean	Percentile
Charlotte Grade 8, Advanced	214	89.50	19	94.48	28	84.17	22	87.42	29	87.92	22	91.96	30	91.99	20	96.16	28
Lake Village Grade 8, Advanced	120	89.45	19	90.78	23	81.41	16	83.25	19	86.43	19	87.66	21	90.52	17	92.02	20
Lorain Grade 8, Advanced	285	91.90	24	94.57	28	85.60	25	88.63	31	88.76	24	92.79	31	93.81	24	97.22	32
Schenectady Grade 8, Advanced	155	92.47	25	94.00	26	81.77	17	87.00	28	87.57	21	91.11	26	91.98	20	95.67	27

NOTE: NR = Number of children for whom Total Reading scores are available for both fall and spring; NM = number of children for whom Total Math scores are available for both fall and spring. Other subtest means and percentiles may be based on larger sample sizes.

Table 3-22

SIGN OF FALL-SPRING CHANGE IN PERCENTILE OF THE MEAN FOR READING SUBTEST, BY SITE AND GRADE

Project	Grade 1	2	3	4	5	6	7	8	9
Catch-Up									
Bloomington	-	-	+	-	+				
Brookport	-	-	+	-	+	+			
Galax	-	-	+	-	+	+			
Providence Forge				-	+	-			
Wayne City	-			-		-	-		
Conquest									
Benton Harbor		+	+	+	+	+			
Cleveland	-	+	+	-	+	+			
Gloversville		-	0	-	+	+			
HIT									
Lexington									
Tutors						-	-	+	+
Tutees						+	-	+	-
Olean									
Tutors							+	+	-
Tutees							-	+	
IRIT									
Bloomington			+	-					
Oklahoma City			+						
Schenectady			+	-					
PTR									
Canton	-								
Dallas	-								
R-3									
Charlotte								+	
Lake Village								+	
Lorain								+	
Schenectady								+	

Note: - = loss in mean percentile fall to spring
+ = gain in mean percentile fall to spring
0 = no change in mean percentile fall to spring.

In Section 3.8 we display a modified norm-referenced analysis that deals with some of the shortcomings of the unmodified analysis.

Perhaps no analysis but the unmodified analysis will satisfy some critics of the PIP concept. We point these critics to the arguments in Section 2, which show that passing the norm-referenced analysis and the criterion of educationally significant growth is not a compelling reason for justification of packaging.

We also point out that with a less conservative approach to the norm-referenced analysis, we would find some evidence for the success of the packaging concept. For example, all third grade IRIT projects showed good gains on the MAT Reading subtest. On the same subtest the Cleveland Conquest project showed fairly large gains over expected growth at the third, fifth, and sixth grades, and Benton Harbor Conquest did well at all grades. Furthermore, when as cautious a procedure as one-tailed confidence intervals confirms that <u>at least</u> normal growth is achieved in almost every case having enough data to decide, we may conclude that packaged projects have the potential for maintaining more than equipercentile growth.

3.8 Special Empirical Studies of the Norm-Referenced Analysis

During the first year of the evaluation of the PIP field test, several peculiar properties of the norm-referenced procedure became apparent. Questions were raised regarding the assumption of equal percentile growth as the basis of the normal growth criterion and regarding the stringency of the criterion for educationally significant growth. The validity of the equal percentile assumption was examined to a limited extent, and the results tended to confirm the "straggler hypothesis," that is, that PIP participants may represent the kinds of students who lose ground over time. Examination of the stringency of the criterion of educational significance indicated that gains necessary to attain educational significance--relative to gains necessary to attain normal growth--increase as a function of grade level.

As a consequence of these findings, SRI conducted a detailed investigation of the properties of the norm-referenced procedure as employed in the PIP evaluation. The examination included not only the equal percentile assumption and the stringency of criteria, but also the statistical properties of the procedure. Properties of interest were the sensitivity of the procedure to the unit of analysis and the effects of the use of the standard deviation of the difference between fall and spring observed scores to approximate the standard deviation of the difference between the spring observed and expected scores.

The detailed results of the study are given in a working paper entitled "A Study of the Norm-Referenced Procedure as Applied to the Evaluation of Project Information Packages" (Kaskowitz and Norwood, 1976). The purpose of this section is to summarize the results of the study and indicate some of the implications for interpreting the results of the norm-referenced analysis.

3.8.1 The Equal Percentile Assumption

The major assumption of the normal growth model that was incorporated in the norm-referenced procedure is that, ceteris paribus, the norm percentile score of a child, class, or site will, on the average, stay the same between pretest and posttest. The sample of children used in the standardization and norming of the MAT was designed to be representative of the entire school population. However, the children included in compensatory education programs, such as the PIPs, are usually different from the entire school population with respect to demographic and socioeconomic characteristics: They are more likely to be members of a minority; they tend to be from lower income families; and they tend to have low pretest scores on standardized achievement tests. More significantly, evidence from the first-year PIP evaluation indicated that they may be representative of students who lose ground over time, relative to the norm population. If this were the case, use of the equal percentile assumption would lead to overestimates of the expected posttest scores used in the test for normal and educationally significant growth. One might then conclude that programs were not effective in raising scores on standardized tests, when in fact they were.

The study of the equal percentile assumption entailed examination of several large-scale data bases containing longitudinal MAT test data on children who would ordinarily qualify for educational programs such as the PIPs. Included were Follow Through (FT) project evaluation data obtained from SRI and Compensatory Reading (CR) program evaluation data obtained from the Educational Testing Service (ETS).* In addition, data from a subset of the MAT norm group were obtained from Psychological Corporation. The norm data consisted of longitudinal information on a large subset of children who were tested both in the fall and spring standardization programs.

*The Follow Through and Compensatory Reading programs are funded by USOE.

From the FT evaluation, data on a subset of children in the comparison group--Non-Follow Through (NFT)--were used in the analysis. These were children who had entered kindergarten in fall 1971 and had been tested in at least two of the subsequent three spring test periods (1973, 1974, or 1975) when the MAT had been administered. From the CR evaluation of children in grades 2, 4, and 6, three groups were of particular interest: children in compensatory reading programs who were participating in the federal school lunch program (CR/SL), children in compensatory reading programs who were not participating in the federal school lunch program (CR/NSL), and children in schools that had no compensatory reading program who were participating in the federal school lunch program (NCR/SL). For individual children, participation in the federal school lunch program was the only available indicator of socioeconomic status. About 75% of the CR/SL children were in schools where compensatory reading programs were funded to some extent by Title I, and about 58% of the CR/NSL children were in schools where the compensatory reading programs were funded to some extent by Title I. These two groups, then, consisted largely of children similar to those for whom the PIPs are targeted.

The subpopulation of the MAT norm group was obtained to examine the relationship between the gains predicted from the cross-sectional standardization design used to derive the percentile norms and the gains empirically obtained from a longitudinal sample. These data were initially analyzed by Dr. Michael Beck of Psychological Corporation, who reported the results in a paper presented at the 1975 Convention of the National Council on Measurement in Education (Beck, 1975).

Two major findings emerged from the examination of the data:

- Expected posttest scores based on the equal percentile assumption tend to be too low for students with extremely low pretest scores.
- Expected posttest scores based on the equal percentile assumption tend to be too high for disadvantaged students, especially disadvantaged minority students whose pretest scores are not extremely low.

The first point is illustrated in Figure 3-1, which is a plot of the empirical growth curve of the CR/SL group on the MAT Total Reading between fall and spring of fourth grade. The empirical growth curve, indicated by the dots in the figure, consists of the posttest mean standard score for each value of the pretest standard score. The solid line represents the relationship between pretest and posttest standard scores under the equal percentile growth assumption. The empirical growth curve for the

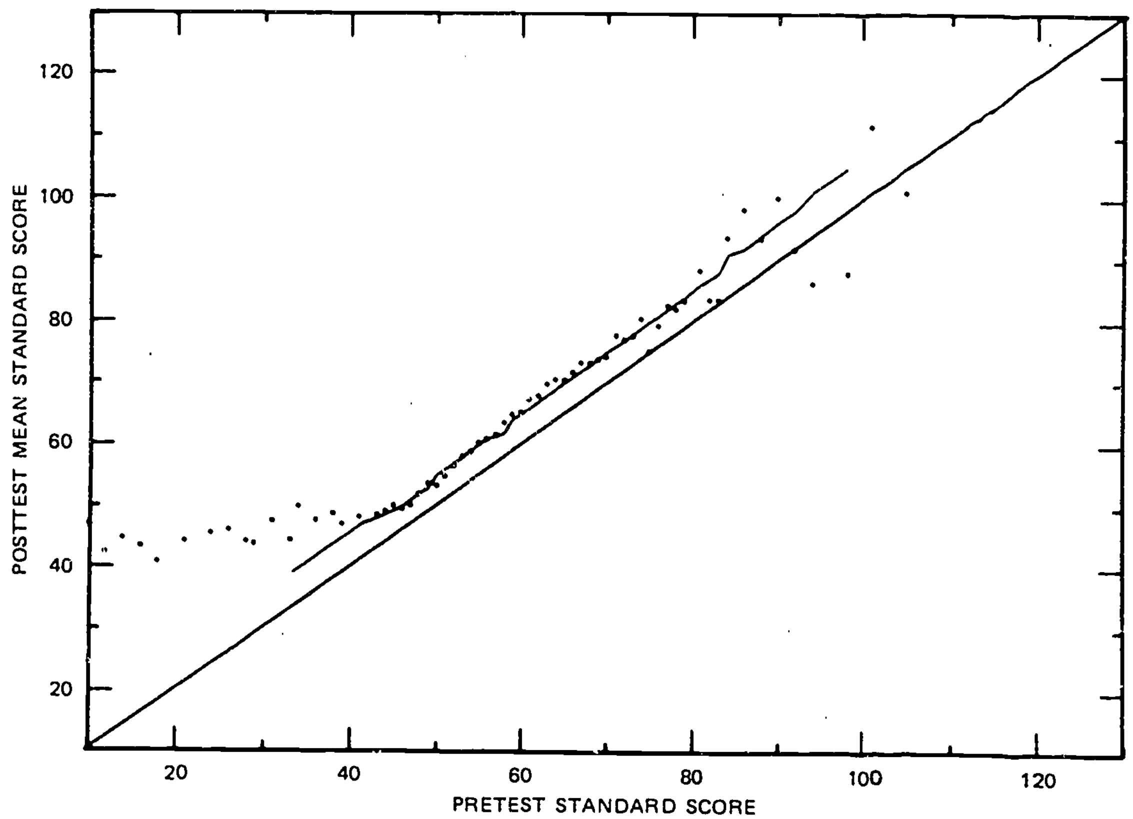


FIGURE 3-1 EMPIRICAL GROWTH CURVE FOR THE FOURTH GRADE CR/SL GROUP ON MAT TOTAL READING: FALL PRETEST, SPRING POSTTEST

CR/SL group tends to track the equal percentile curve in the midrange of fall scores, between standard scores of 46 (the 6th percentile) and 88 (the 94th percentile). However, the equal percentile curve shows that students scoring below the 6th percentile in the fall tend to exhibit gains that are greater than expected in the spring.

The second point, regarding gains for disadvantaged children, is illustrated in Figures 3-2 and 3-3. Figure 3-2 shows the empirical growth curve for the NFT group on Total Reading between the spring of first grade and the spring of second grade. Again, the equal percentile model underpredicts the gains made by pupils with extremely low pretest scores. However, it also overpredicts gains for children with pretest standard scores above the extremely low level, as indicated by the empirical growth curve lying below the equal percentile curve for pretest standard scores above 25. Even with the phenomenon of higher than expected gains for children with low pretest scores, the NFT group had an average drop in percentile rank of 6.9 points between the spring of first and second grade and 2.4 points between the spring of second and third grade on Total Reading. For Total Math the average drop was 6.3 percentile points between the spring of first and second grade, but no drop between second and third grade.

The CR evaluation groups showed gains in their percentile ranks on Total Reading between fall and spring across the three grade levels included in the evaluation. The gains averaged about 10 points for second grade, between 3 and 5 points for fourth grade, and about 2 points for sixth grade.

The dramatic decline in percentile ranks for the NFT group over the two-year period supports the straggler hypothesis--that is, in the absence of intervention programs, children targeted for compensatory education programs for the disadvantaged will lose ground relative to the norm group. The data from the CR evaluation, on the other hand, are not unequivocal in denying the straggler hypothesis. For one thing, children in the CR/SL and CR/NSL groups were participating in compensatory reading programs that in fact may have reversed a decline in percentiles. The NCR/SL group that appeared to serve the role of comparison group had percentile gains that equaled or exceeded those of the two CR groups. Furthermore, the NCR/SL group has characteristics that would call into question the assumption that these children are typical of those who would be in compensatory programs. They are disproportionately from the South, from moderate-size cities, and of nonminority status. Finally, none of the NCR/SL students, as contrasted with 44% of the CR/SL students, were in schools where the estimated percent of students from families receiving public assistance exceeded 25% of the school population.

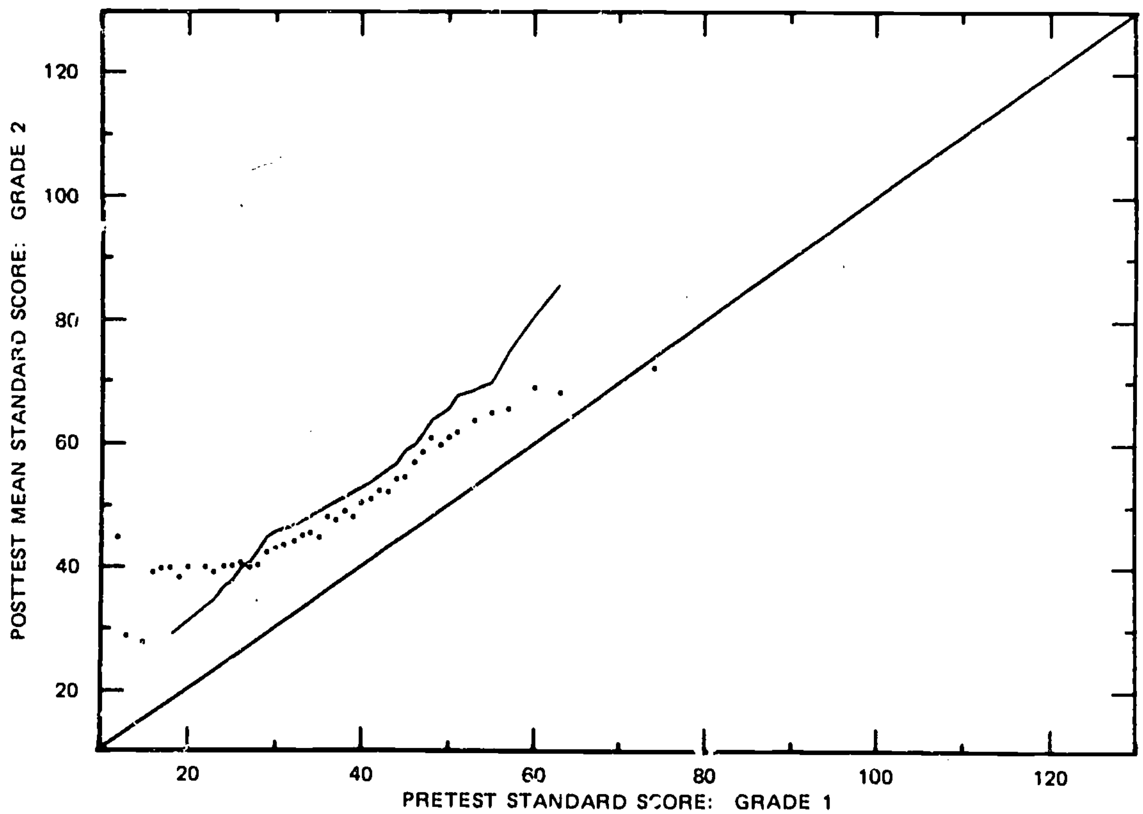


FIGURE 3-2 EMPIRICAL GROWTH CURVE OF THE NFT GROUP, FIRST TO SECOND GRADE, ON MAT TOTAL READING: SPRING PRETEST AND POSTTEST

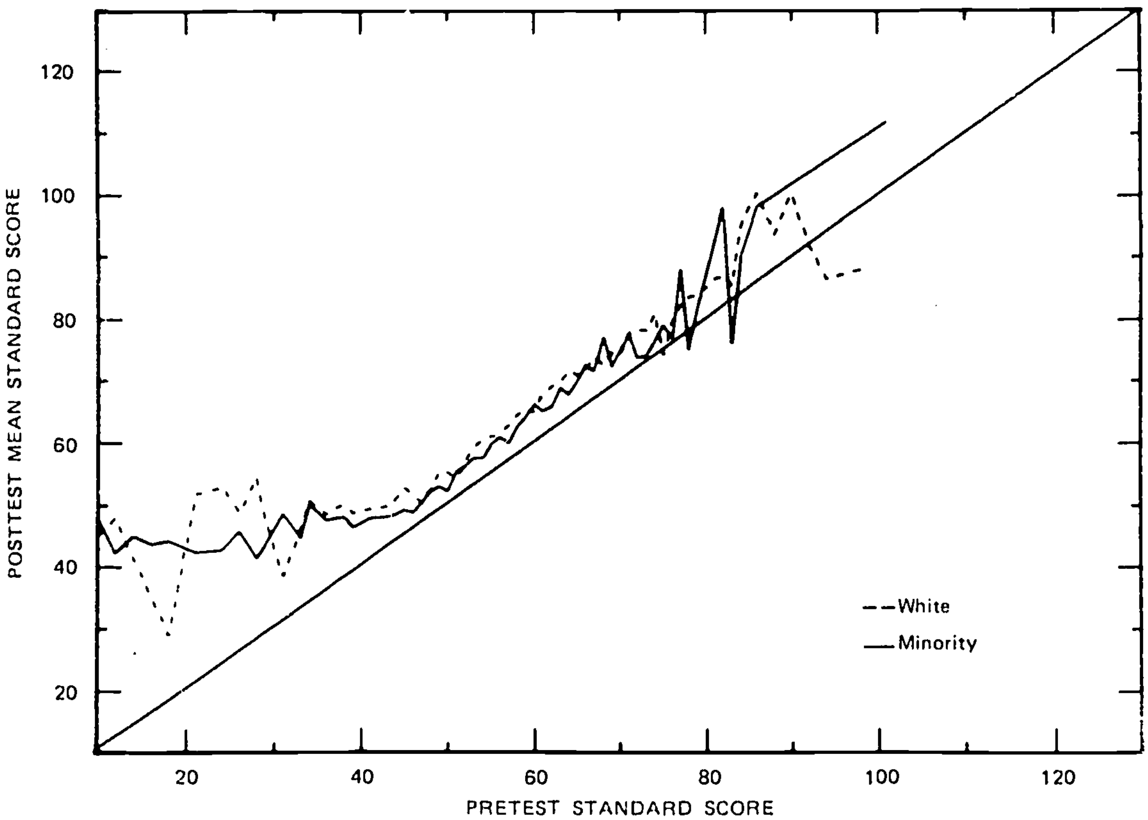


FIGURE 3-3 EMPIRICAL GROWTH CURVE FOR THE FOURTH GRADE CR/SL GROUP ON MAT TOTAL READING, BY MINORITY STATUS: FALL PRETEST, SPRING POSTTEST

It would appear that out-of-level testing probably does not explain the extraordinary gains in second grade. Pelavin and Barker (1976) in fact found that when pupils are tested within a short time period on both the Primary I and Primary II Reading subtests that standard scores on Primary II tend to be higher than standard scores on Primary I. This would mean that the out-of-level testing in the spring may have suppressed the gains in standard scores on Total Reading.

Figure 3-3 is illustrative of the relative performance of white and minority pupils. It shows the empirical growth curves for children in the fourth grade CR/SL group on Total Reading; white children are indicated by the broken line and minority children by the solid line. Between the pretest standard scores of 40 (the 2nd percentile) and 65 (the 48th percentile), the empirical growth curve for minority students is consistently below that for white students. For the NFT group and each CR group, this was a consistent pattern for Total Reading and Total Math: Minority students consistently gained less in percentile ranks (or lost more) than did white students. Of course, minority status was confounded with other demographic and socioeconomic characteristics, such as median income, size of place, and region, so that minority status of itself cannot be identified as the sole factor in accounting for differences in growth. Nevertheless, since so many of the disadvantaged children targeted for programs such as the PIPs are of minority status, the differences in performance are of particular interest and import.

3.8.2 The Statistical Properties of the Procedure

Two questions were examined regarding the statistical properties of the norm-referenced procedure: (1) How sensitive is the test to the unit of analysis? and (2) How good an approximation is the standard deviation of the difference between the pretest and posttest scores to the standard deviation of the difference between the posttest and expected posttest scores?

Algebraically, the t statistic used to test for normal growth may be expressed as follows:

$$t = \frac{\bar{X}_{post} - f\left(\bar{X}_{pre}\right)}{\sqrt{\frac{V_{post - pre}}{N}}} ,$$

where

$\bar{X}_{post}$ = mean standard score on the posttest

$\bar{X}_{pre}$ = mean standard score on the pretest

$V_{post-pre}$ = sample variance of the difference between pre- and post-test standard scores

N = number of cases

f(x) = function that gives the "expected" posttest score, given the pretest score.

The current procedure converts the pretest mean score to a predicted posttest mean score. The conversion could also be made individually for each student's score; the individual predicted scores could then be aggregated to yield the average posttest predicted score. Since the function that converts standard scores to percentiles is intended for use at the individual student level, the question arises whether this alternative might yield different results. If the transform, f, is linear, there would be no difference between the two procedures. In the criterion study, it was found that under the equal percentile assumption the function could be closely approximated by a linear function so that the current procedure is relatively insensitive to the unit of analysis. However, under alternative nonlinear models of normal growth, different results can be obtained, depending on the unit of analysis selected.

The question of the variance estimate used in the denominator was discussed in Section 2.3.3. Horst, Tallmadge, and Wood (1975) recommend using $V_{post-pre}$ to estimate $V_{post-\widehat{post}}$, where $\widehat{post}$ is an estimate of the predicted posttest score under the normal growth assumption. The relationship between these two variances depends on the form of the function that predicts the posttest score from the pretest score and the relationship between pre- and post-test scores. As indicated earlier, under the equal percentile assumption, the function f is approximately linear. Since the coefficient of the first-order term is close to one, the variance of the difference between the pre- and post-tests would appear to be an adequate approximation.

Four sets of data on Project Catch-Up from the first-year PIP evaluation were reanalyzed to assess the impact of modifications in the statistical procedure on the results of the norm-referenced analysis. The transformation from fall score to expected spring score was applied at the student level rather than at the site level. The mean and the variance of the difference between the student observed and expected spring scores were used in the calculation of the t statistic. The results are summaried in Table 3-23.

Table 3-23

COMPARISON OF RESULTS OF ORIGINAL AND MODIFIED NORM-REFERENCED PROCEDURE

Test and Procedure	Number of Students	Gain over Expected Gain	SD*	t Test	Meets Normal Growth
Total Reading (grade 3)	18				
Original procedure		1.60	4.33	1.56	Unknown
Modified procedure		0.94	5.04	0.79	Unknown
Total Math (grade 5)	22				
Original procedure		7.14	5.67	5.91	Yes
Modified procedure		3.15	5.04	2.93	Yes
Total Reading (grade 5)	19				
Original procedure		1.47	4.13	1.55	Unknown
Modified procedure		2.05	4.70	1.90	Unknown
Total Math (grade 6)	27				
Original procedure		1.11	6.77	0.85	Unknown
Modified procedure		1.06	6.70	0.82	Unknown

*For original procedure, $SD_{post-pre}$; for modified procedure, $SD_{post-\widehat{post}}$.

For three of the four sets of data shown in Table 3-23, the numerator of the t statistic, gain over expected gain, is lower under the modified procedure than under the original procedure; the denominator tends to be the same under either procedure. As a result, the t values tend to be smaller under the modified procedure. In all of these cases, the conclusion regarding normal growth would have been the same under either procedure. However, in situations where normal growth is only narrowly achieved under the original procedure, it may not be achieved under the modified procedure.

3.8.3 Stringency of the Criteria

Ideally, the criteria of normal growth and of educationally significant growth should be established and implemented so that the degree of difficulty in attaining the criteria is independent of factors such as grade or pretest score. A procedure by which it is easier to demonstrate program effectiveness at one grade level than at another would be unfair.

The stringency or the difficulty of meeting a criterion can be judged both in educational and statistical terms. Educationally, the difficulty of meeting the norm-referenced criteria is related to the effort necessary for achieving the necessary gains, where effort is measured in terms of educational resources and time needed for students to acquire the necessary skills. While the standard score metric is purported to be an equal interval scale of achievement, it is probably not the case that a specified gain in standard score requires the same amount of effort independent of grade level or initial standard score. Also, as was emphasized earlier, effort per se is not enough; it must be effort directed to the acquisition of skills measured by the MAT.

The equal percentile assumption for normal growth demonstrates the differences in standard score gains across grade levels necessary for normal growth. For Total Reading, a gain of between 7 and 9 standard score points is necessary to maintain normal growth at the second grade between fall and spring. This decreases to a gain of between 1 and 4 standard score points for eighth grade. In a few instances, in the upper grade levels, the equal percentile assumption dictates that zero gain between fall and spring is sufficient for normal growth. Below the 50th percentile, the specified standard score gains necessary for normal growth are quite uniform across percentile ranks within grade.

The substantial differences across grade may be attributed to any of a number of factors:

- The test items become increasingly irrelevant to the types of skills being taught in the upper grades.
- Students tend to reach an asymptote in their acquisition of reading skills, and additional gains require much more effort than at the lower grade levels.
- Less time and effort are spent in the upper grades in acquiring the skills measured by the MAT.
- The standardization procedure was defective.

If it is assumed that the standardization program produced valid norms, then either students appear to reach some asymptote in the upper grade levels on reading achievement or the curriculum at the upper grade level is irrelevant to the skills measured by the MAT. If the curriculum is irrelevant, it may not be the case that the observed low gains in standard scores indicate a great degree of difficulty in achieving growth. It may merely mean that students are not spending much time learning skills relevant to items on the MAT.

The difficulty of achieving specified gains on the MAT cannot be assessed with any degree of accuracy, given current educational theory. However, the gains expected under the normal growth assumption may be taken as a baseline for assessing the stringency of the criterion of educational significance. That is, the gains by the MAT standardization group may be taken as representative of the output of programs with an average effort expended at each grade level.

This was the point of view adopted in Section 2.3. It was found there that, since the rate of change in the standard deviation of standard scores was small, the one-third norm standard deviation criterion for growth is constant over grade levels. Therefore, with the decrease in the expected normal growth across grade levels, it appears that the criterion of educationally significant growth becomes increasingly stringent as grade level increases. That is, greater effort seems necessary at the upper grade levels, relative to the effort normally expended at those levels, to obtain gains that meet the criterion of educational significance. The one-third standard deviation criterion for educationally significant growth was proposed by Horst, Tallmadge, and Wood merely as a rule of thumb. The results of the study of stringency reinforce this point of view: No empirical or theoretical basis exists for selecting the one-third criterion, and at least from one point of view this criterion becomes increasingly difficult to meet as a function of grade level.

Statistically, the issue of stringency appears to be more clear-cut, and one factor--the number of students in the analysis--appears to be more critical than the grade level or pretest scores. From a statistical point of view, the stringency of the criterion may be expressed in terms of the power function that describes the chance of meeting the criterion given particular gains (under the ideal conditions underlying the application of the t test). For the norm-referenced procedure suggested by Horst, Tallmadge, and Wood, the power function may be expressed as:

$$P\left(T \geq t_{0.025,\ N-1} \middle| \delta\right) \quad ,$$

where T is the test statistic described in Section 2; $t_{0.025,\ N-1}$ is the 0.025 critical point of a student's t distribution, with N-1 degrees of freedom; and δ is the noncentrality parameter expressed as $\Delta\sqrt{n}/\sigma$, where Δ is the true difference between the observed and expected spring standard scores and σ is the standard deviation of these differences. Figure 3-4 shows the power curves for a few selected sample sizes. Note that, in all cases, the probability of passing the normal growth criterion--given that Δ is zero--is 0.025. That is, if the "population" gains in standard score are exactly what is expected under the normal growth assumption, the normal growth criterion will be passed in only 25 out of 1000 replications. Obviously, for a given value of Δ/σ, the chances of meeting the criterion increase as the number of students in the analysis increases. For example, for $\Delta/\sigma = 0.3$, the chance of meeting the criterion is greater than 8 out of 10 when the number of students is about 100, drops to less than 4 out of 10 when the number of students is about 30, and is less than 2 out of 10 when the number of students is about 10.

From a statistical point of view, it is plausible and reasonable, of course, to have a more stringent requirement for normal growth as the sample size decreases. However, in most field evaluations the number of children in the program to be evaluated is not under the control of the evaluators. Therefore, the stringency of the criterion depends to some extent on extraneous factors such as the number of sites where the program was implemented, the number of children at the sites who qualified for the program, and the optimum number of children that could be accommodated in the operational design of the program. For example, some PIPs have as few as 4 students at a given grade level and others have as many as 779. Other things being equal, including the actual impact of a program on achievement, the program with the larger number of children has a much greater chance of demonstrating its effectiveness.

3.8.4 Modifications of the Norm-Referenced Procedure

In the preceding discussion, potential weaknesses of the norm-referenced procedure were described. These include:

- On some tests, the expected posttest standard score based on the equal percentile assumption is too low for students with extremely low pretest scores.
- There are indications that the expected posttest standard scores are too high for disadvantaged students, especially disadvantaged minority students.

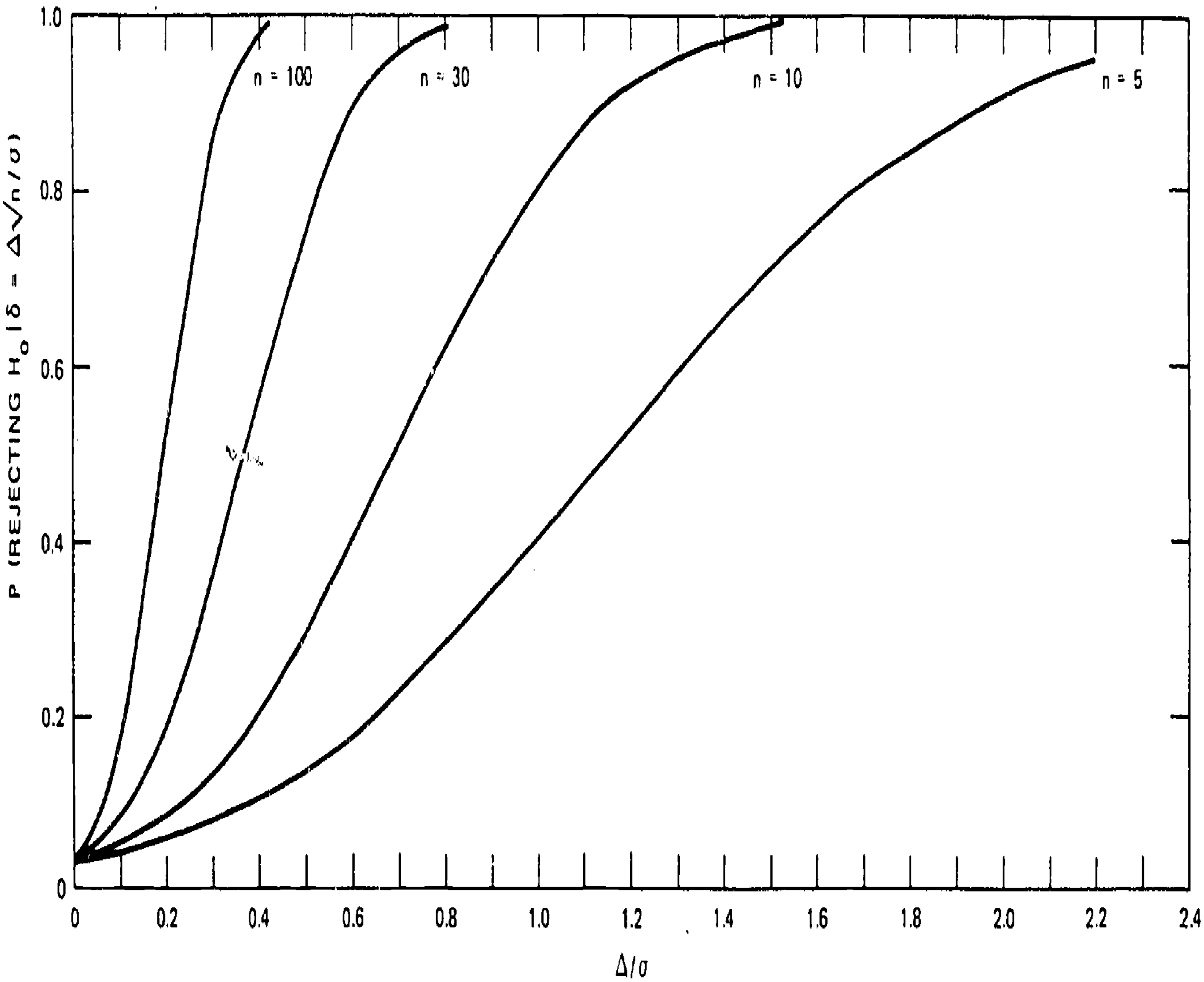


FIGURE 3-4 POWER CURVES FOR SELECTED SAMPLE SIZES, α = 0.025

- The criterion for educationally significant growth may not be of uniform stringency across grade levels.
- The stringency of the criteria for normal growth and for educationally significant growth depend on the number of students in the evaluation.

This list does not touch upon some of the more basic criticisms of the norm-referenced procedure that were discussed in Section 2. If an evaluator agrees that the conceptual basis for the norm-referenced analysis is extremely weak, no minor modifications of the procedure will be satisfactory.

The modifications necessary for satisfying the first two points above are relatively straightforward: Raise the expected posttest standard score for extremely low pretest scores; otherwise, lower the expected spring standard score. Furthermore, use different expected posttest scores for students of different minority status. It is not at all clear what modifications of the current procedure can be made to make stringency uniform across grades. For one thing, there is no educational theory to quantify the effort necessary for achieving a specified gain. Therefore, assessing the stringency of the normal growth criterion is not possible. If the normal growth criterion is taken as a baseline, under the assumption that it represents the results of an average level of effort, it is not clear how educationally significant growth should relate to normal growth. For example, if the standard score gain necessary for exhibiting educationally significant growth is taken as a constant multiple of the standard score gain necessary for normal growth, it is not at all clear what multiple should be used. The current procedure provides no guidance because the one-third standard deviation criterion was offered as a rule of thumb with no educational basis for its adoption. Under these circumstances, no modification of the educationally significant growth criterion was examined.

The dependence of the stringency of the criterion on the number of students in the evaluation is also a factor that is not amenable to a modification of the procedure after the data have been collected. Theoretically, part of the design of the evaluation should be a specification of the number of children to be included, where this number is based on considerations of the power of the statistical procedure. When the number of children included in an evaluation is determined in a haphazard fashion, there does not appear to be any post hoc formal procedure for making the statistical analysis independent of the number of children.

In this concluding section, then, we will examine the impact on the norm-referenced analysis of changes in the specification of normal growth for fourth grade Total Reading. Fourth grade was selected because it was the only grade in which the same battery level of the MAT had been administered to the MAT standardization group, the compensatory reading evaluation group, and the PIP evaluation group. The CR/SL group was used to derive the function describing the expected posttest score, given the pretest score. A separate function was derived for each of three ethnic/racial groups: whites, blacks, and Spanish surname. A regression analysis was used to derive the function in a pretest standard score range between 46 and 88, corresponding to a percentile range between 6 and 94. The summary statistics are presented in Table 3-24.

Table 3-24

SUMMARY STATISTICS FOR ESTIMATION OF EXPECTED POSTTEST TOTAL READING STANDARD SCORE, GIVEN PRETEST STANDARD SCORE

Race/Ethnicity	Number of Students	Intercept	Slope	r^2	SD	SE Slope
White	1238	3.32	1.04	0.69	6.2	0.02
Black	619	-1.60	1.11	0.57	5.8	0.04
Spanish surname	179	8.31	0.93	0.63	5.8	0.05

Note: Statistics were derived from data provided by ETS.

It was felt that the functions still over-predicted posttest standard scores because the time interval between pre- and post-test may have been as much as a month longer than the interval for the PIP evaluation. Also, these children had been in some form of compensatory reading program. It is not clear whether children who were in the PIP program would have been in other compensatory programs if the PIPs had not been implemented. If the children had been in other compensatory programs, such as those funded under Title I, the CR/SL group would appear to be a good comparison group. Otherwise, some adjustment to the functions would be necessary to account for the gains attributable to the compensatory reading program. For current purposes, each intercept was decreased by 1.5 standard score points to represent the potential effect of about a month's increase in the time interval and the potential effect of the compensatory reading program. Pelavin and Barker (1976) proposed a rate of growth of about 0.75 standard score points per month as a rule of thumb in their study of the MAT. Most studies of compensatory education programs have found only small effects attributable to the program. Therefore, a cumulative effect of about 1.5 standard score points appears to be reasonable for this study.

For students below the 6th percentile lower bound, a constant expected spring score was postulated for each racial/ethnic group. Prior results (see Figure 3-1) had indicated that the expected spring score for a student with fall standard scores below the 6th percentile was approximately independent of the particular pretest score. The expected posttest score was found to be approximately equal to the regression lines evaluated at a pretest standard score of 45. A similar procedure was used for students with scores about the 96th percentile.

The student was used as the unit of analysis. Each student's fall score, F, was used to derive an expected spring score, E. The difference, D = S - E, between the observed spring standard scores, S, and the expected spring scores, E, was then calculated. These differences were used to calculate the t statistic:

$$T = \frac{\bar{D}}{SD/\sqrt{n}}$$

where

$\bar{D}$ = mean difference

SD = standard deviation of the differences

n = number of students.

Under the assumption that the deviations between observed and expected scores are approximately independent and normally distibuted, T will have a student's t distribution with n-1 degrees of freedom.

The same criterion as that used in the unmodified analysis was used to test for normal growth: If the 95% confidence interval on the mean gain was totally above zero, the normal growth criterion was said to be attained; if the confidence interval was totally below zero, normal growth was said not to be attained; otherwise, it was assumed unknown whether or not normal growth had been attained.

Table [illegible] gives the results of the original and modified analysis with respect to the normal growth criterion for fourth grade Total Reading. Overall, no dramatic shift in conclusions is evident. One project, Conquest, changed from "unknown" to "yes" relative to meeting the normal

Table 3-25

RESULTS OF ORIGINAL AND MODIFIED NORM-REFERENCED
PROCEDURE FOR FOURTH GRADE TOTAL READING, BY PIP AND SITE

PIP/Site	Number of Students	Gain Over Fall*	Original Analysis: Gain Over Expected*	Original Analysis: t Test	Original Analysis: Meets Normal Growth	Modified Analysis: Gain Over Expected*	Modified Analysis: t Test	Modified Analysis: Meets Normal Growth
Catch-up	83	2.96	-2.04	-3.46	No	-1.48	-2.81	No
Bloomington	40	2.73	-2.28	-2.70	No	-1.75	-2.23	No
Brookport	8	3.75	-1.25	-0.75	Unknown	0.01	0.01	Unknown
Galax	7	4.43	0.43	0.11	Unknown	-2.49	-0.95	Unknown
Providence Forge	20	3.05	-1.95	-2.08	Unknown	-0.94	-0.91	Unknown
Wayne City	8	1.88	-2.25	-2.64	No	-2.11	-2.36	Unknown
Conquest	108	5.56	1.08	1.98	Unknown	1.24	2.46	Yes
Benton Harbor	28	7.36	3.19	2.75	Yes	3.45	3.53	Yes
Celevland	53	4.77	0.49	0.65	Unknown	0.80	1.17	Unknown
Gloversville	27	5.22	0.22	0.22	Unknown	-0.18	-0.17	Unknown
IRIT	34	3.44	-1.23	-1.39	Unknown	-1.18	-1.58	Unknown
Bloomington	28	4.39	0.12	0.12	Unknown	-0.46	-0.56	Unknown
Schenectady	6	-1.00	-6.00	-6.71	No	-4.56	-4.94	No

*In standard score units.

growth assumption. Only one site, Wayne City, changed from "no" to "unknown." Under the modified procedure, the gains over expected increased for the PIPs shown in the table. At a few sites, the gains over expected decreased because of the increase in expected spring scores for students with extremely low fall scores. The difference between the gains of students with extremely low pretest scores and students with higher scores is revealing: For the 44 pupils who scored below the 6th percentile in the fall, the average gain in standard score was 8.4 points; for the 181 pupils who scored at the 6th percentile or above, the average gain was only 3.3 standard score points.

Based on the above results, it would appear that modest modifications of the norm-referenced procedure will not lead to different conclusions regarding the effectiveness of the PIPs, except in marginal cases. In the remaining sections, results of the curriculum-referenced analysis are reported.

4 FIELD EVALUATION OF IMPLEMENTATION

4.1 Introduction

Results of field activities undertaken to satisfy our principle of description are discussed in this section of the report. In accordance with the assumption that, all other things being present, MAT items will be learned from an appropriate curriculum, given a reasonable length of exposure with a competent teacher, we attempted to uncover exactly what the PIP-specified instructional techniques and curricula were. We also desired to assess the impact that the degree of implementation of the instructional component had on test scores.

In this section we discuss the PIP instructional component and how we assessed the degree to which it was implemented. Because of the PIPs' emphasis on management, last year we examined the management features of the PIP projects. This year we were more interested in the hardware/software component of the PIPs and in the classroom instructional procedures, which were scattered throughout the packages.

We immediately discovered that very little was specified relative to project instructional style. The PIPs assumed that qualified project staff already know how to teach in the style required. The PIPs also emphasized the use of widely available published materials. The immediate implication was that the PIPs would not necessarily place an innovative instructional process into a conventional atmosphere. At this early stage of our investigation, for all we knew, the PIPs' curricula and styles would be substantially the same as those they were intended to supplement,* or in the case of R-3, to replace. In any event, the PIPs' reliance on published materials (except for R-3) made it quite likely that, insofar as test scores were concerned, PIP effects and school effects would be the same. It would be just as reasonable to attribute gains to the regular school as to the PIP-induced add-on, if the two teaching styles and curricula were substantially the same.

---

* The IRIT and Catch-Up PIPs warn that their projects must be made distinct from the regular curriculum.

We set up our site visit procedures to accomplish two objectives. The first was to assess implementation and the factors that influenced it, which would allow us to make reasoned policy recommendations about packaging. A good assessment of implementation would free us from complete dependence on the norm-referenced analysis for evidence of project success. If we could show that test scores increased with good implementation, we would have evidence consistent with PIP effectiveness.

The second objective for our fieldwork was to collect information on the curriculum that the project teacher actually used. These data would make it possible to compare PIP-specified curricula with the MAT items and to compare MAT items with the items in the tests used for Dissemination and Review Panel approval.

We had some success in achieving both objectives. Section 5 presents results of our curriculum analysis and describes the data we have for comparing the MAT with the original Dissemination and Review Panel test. Section 6 contains formal analyses that use our data on the curriculum and on implementation to assess achievement impacts.

Section 4.2 describes our method of approach for recording on-site observations.

4.2 Methodology for On-Site Observations

An evaluation that attempts to assess the worth of an innovation without systematically describing the innovation, as realized in the field, is in danger of evaluating the wrong program. The program in the field is, after all, the presumed causal agent for any outcomes of interest.

The methodological problem is how an observer may correctly describe what is in the field and how he communicates this to the interested audience. Broadly speaking, we distinguished two paths as being available to us. The first, we call the objective-subjective, or hard-nosed approach. The second, we call the subjective-objective, or the soft-headed approach.

In the objective-subjective approach, the observer's notebooks are forms in which he records the presence of objective factors, such as the lighting and seating patterns in a classroom. There may be space to record some fairly molar behaviors, such as a teacher asking questions. Characteristically, in the hard-nosed method data are recorded at a fairly fine level of detail, scored, and then statistically aggregated into measures of various factors. The analyst may even construct a variable called "implementation."

In the subjective-objective approach, the observer takes his notes in ordinary language and reports them novelistically. The soft-headed approach is characteristically conducted at a molar level, with judgments and conclusions being stated and justified in text.

We call the hard-nosed approach objective-subjective because it is fairly objective during data collection but subjective at analysis time, since the choice of metric is subjective. On the other hand, we call the soft-headed approach subjective-objective because at data collection time the observer makes judgments, but at analysis time the analyst can only treat the judgment as objective. Characteristically, the objective-subjective approach ends in the analysis of a real-valued implementation scale, while the subjective-objective method ends in the analysis of the yes/no judgment of implementation. Much can be said for, and against, each approach. Last year we relied principally on the objective-subjective technique. This year we relied principally on the subjective-objective technique.

The first- and second-year observation methodologies are related in that, in both years, we used forms as the primary components of the observers' notebooks and began the second-year site visit by using the first year's haphazardly selected sample of schools and teachers. However, our intention was not to repeat last year's general assessment of implementation. Rather, we wished to focus in depth on the extremes of implementation; we used the information from the first year to select those extremes. Ultimately we hoped to determine whether the successful implementation of the PIP instructional program made a difference in the children's achievement scores.

Table 4-1 shows the dates of the site visits, by project, and the site visitor responsible for each. Generally the visits were made after fall testing and before spring testing. Only two PIPs had more than one site visitor. Because it was not economically possible to send more than one visitor to each site, the validity of our assessments of teacher implementation is principally dependent on the judgment and objectivity of a single site visitor. Although undetected errors are possible, we feel that we have been successful in comprehending each PIP's intent.

The steps we took to define the intent of each PIP and the resultant interpretation of intent are described below.

Table 4-1

SITE VISITS FOR OBSERVATION AND INTERVIEWS

Project	Visit Date Fall 1975	Visit Date Spring 1976	Site Visitor
Catch-Up			
Bloomington	12/9-12/12	2/18-2/19	Margaret Needels
Brookport	12/4-12/5	3/11-3/12	Phil Giesen
Galax	12/10-12/12	3/18-3/19	Phil Giesen
Providence Forge	12/9-12/10	3/15-3/16	Phil Giesen
Wayne City	12/3-12/4	3/9-3/10	Phil Giesen
Conquest			
Benton Harbor	12/12-12/15	3/11-3/15	Casse Duarte
Cleveland	12/8-12/10	3/8-3/10	Casse Duarte
Gloversville	12/16-12/18	3/16-3/19	Casse Duarte
HIT			
Lexington	12/7-12/9	3/10-3/12	Dorothy Booth
Olean	12/2-12/5	3/17-3/19	Dorothy Booth
IRIT			
Bloomington	12/11-12/12	2/16-2/17	Margaret Needels
Oklahoma City	12/15-12/16	2/9-2/10	Margaret Needels
Schenectady	No visit*	2/12-2/13	Margaret Needels
PTR			
Canton	12/1-12/2	4/12-4/15	Jay Cross
Dallas	12/3-12/5	3/31, 4/1, 4/7	Jay Cross
R-3			
Charlotte	12/10-12/12	3/17-3/19	Georgia Gillis
Lake Village	12/11-12/12	3/22-3/24	Dorothy Booth
Lorain	12/5-12/9	4/27-4/29	Georgia Gillis
Schenectady	12/2-12/4	3/9-3/11	Georgia Gillis

*No site visit because of teachers' strike at this site.

4.3 Identification of PIP Instructional Components

4.3.1 Results of the PIP Washington, D.C., Conference

At the end of the first year's study, we were aware that most of the tryout projects had been generally successful in implementing the explicit PIP specifications. These were the management specifications and those areas of the instructional program that were clear enough to make verification possible. However, problems at each project caused dissimilarities across sites within the same PIP. With the increased emphasis on achievement gains in the second year, it was important that the projects be given an opportunity to resolve problems caused by attempts to meet certain specifications literally and to discuss whether they were obliged to comply with specifications in cases where they had already established nonreplications. Thus, in September 1975, USOE held a two-day conference in Washington for project directors and selected teachers and administrators from each project site. Representatives from the developer sites, RMC, and SRI also attended. Because the purpose of the conference was to promote uniformity in interpretation and procedure among the tryout sites within each PIP, project directors and teachers exchanged experiences and, with the help of the originating site director, agreed on resolutions of some of the problems they had encountered in trying to follow the PIP specifications. Records were kept to provide early information for PIP revisions.

Following the conference, summaries of the agreed-upon interpretations, revisions, and clarifications for each PIP were printed and distributed to the project directors. The complete summaries are reproduced in Appendix C. The following examples show the substance of the changes.

- Catch-Up--Criterion-referenced tests were interpreted as teaching tools to be used about once a week so that teachers could keep abreast of an individual student's needs. Teachers were cautioned that the use of Catch-Up materials in students' regular classrooms would minimize the materials' effectiveness in the project; yet, students' understanding would be increased by the use of project materials that correlate with those used in the regular classroom.

- Conquest--The role of the supervising clinician in the Conquest program was clarified by setting down some specifications for that person's duties, hours, and teaching assignments. Specifically, the supervising clinician's main duties were to assist the project director with training and administration, to monitor and assist clinicians, and to teach some students, with the amount of time spent on each of these duties to vary with the size of the project and the number of clinicians to be observed.

A revised diagnostic test battery was issued to participants with a warning that the 12 tests must be administered in the order of listing. However, it was agreed that the diagnostic process itself is more important than the tests used and that some substitutes might be acceptable.

- HIT--Many clarifications of the instructional procedures resulted from the HIT meetings. Drill in math uses many different materials, commercial or teacher-made, designed to teach basic facts; there should be sufficient materials to keep drill from being unnecessarily monotonous. Drill in reading is aimed at fluent word recognition, which should not be interrupted by having students spell or write words letter by letter. Drill should be varied by occasionally making a game or contest out of drill words.

- IRIT--IRIT participants clarified some instructional and management specifications: Basal readers that are incorporated in the program at the request of sending teachers should supplement but not replace core materials; learning machines could be used for practice or as motivators but should not be given excessive emphasis; IRIT teachers should give students end-of-cycle placement tests and advise sending teachers of the results; teachers should specialize in one area of instruction rather than rotate among areas during a given year.

- PTR--Tryout project directors insisted that materials be included in the PIP for use with students who do not have reading readiness skills. The Alphabet Skills Book, with which they had achieved some success, was recommended for this purpose. The originating site's project director reported that the basal reading series does not have to match the tutoring kits used in PTR and urged that future PIP projects be allowed to adapt their programs to local conditions. All participants agreed that future projects should have access to the experienced project directors for technical assistance with management and instructional problems.

- R-3--The R-3 participants were reminded that gaming/simulations (G/S) and contracts were integral parts of the program and were urged to use them more frequently. Summer workshops were recommended to enable teachers to adapt the G/S activities for easier integration into the curriculum.

As a result of the Washington conference, all participants had a better idea of how to implement some parts of their programs and of the importance to the evaluation of replicating the PIP guidelines. There were better understanding and greater consensus about which PIP specifications were rigid and which flexible, and about which were of very high priority and which of lower priority in claiming replication.

In our efforts to systematize the description of the PIP instructional programs, we found it necessary to provide a framework from which we could view all of the specified classroom processes. Following the Washington conference, detailed program descriptions for each PIP were written to draw together suggestions and specifications from those sections of the PIP components that bore on instruction and to incorporate the revisions and clarifications formulated at the conference. The following section reports the results of our efforts. We regard these descriptions as the criterial interpretation of the PIPs' instructional components as modified by the Washington conference.

4.3.2 PIP Instructional Program Descriptions

The instructional programs were not presented intact in the PIPs; rather, some elements of instruction were described briefly in eight of the nine components of each PIP. Since we desired to know how all of these pieces fit into a coherent approach, we wrote a description of each PIP's instructional program. This effort entailed pulling out the appropriate sections from the PIPs and incorporating the revisions and clarifications that resulted from the discussions at the Washington conference.

While the descriptions were being written, it became clear that we were flavoring them with our own interpretations of what the guidelines should convey, because our understanding of the programs had changed since our first reading of the PIPs. Our visits to some of the originating sites, our discussions of interpretations throughout the first year with the RMC specialists and with the project staff at the tryout sites, and our own educational philosophy had all contributed to a clearer idea of what each PIP program should be. The revised program descriptions for each PIP follow. It is against these descriptions that we judged implementation and the quality of teaching at each project site.

4.3.2.1 Catch-Up Instructional Program

Project Catch-Up is designed to help children who are in the lower quartiles of their age-mates in reading and math skills. It is not intended, however, to include children who have serious emotional problems or who suffer mental retardation. The idea is to focus attention on those children who, somewhere along the line, have failed to learn the normal content materials (whether prereading, math skills, or materials with more advanced content).

The causes for these deficiencies in content knowledge were never analyzed in depth at the original site, but were assumed to be typical causes--for example, less than sufficient attention at home, lack of exposure to a stimulating environment, or, perhaps, language difficulties.

The primary philosophic belief underlying the Catch-Up style is that any normal child can learn at the same rate (as his age-mates) as long as this belief is conveyed to the child and that he will achieve success through exposure to a planned instructional program. The key to the program, therefore, is a coordinated effort that focuses on specific needs of the child and is carried out in a positive "you can learn" atmosphere. The essential elements of the PIP instructional program are:

- Identification of the specific content deficiencies.
- Assignment of relevant materials.
- Maintenance of a positive atmosphere, both by assuring success and by giving positive feedback.
- Maintenance of the child's enthusiasm.

Specific content deficiencies in reading are identified primarily through the use of criterion-referenced evaluation materials, specifically Random House. Problem areas, as well as the component parts of these areas, are identified through this criterion-referenced series, and skill needs are identified in detail. For example, "identifying final single consonant sounds" and "identifying vowels modified by r," are skills that would be examined.

The means for identification of problem areas in mathematics are less well defined by the PIP. The main vehicle appears to be the placement exam in the Sullivan Math Program, but this exam does not refer to any specific technique or system for identifying weaknesses in math.

In both reading and math these diagnostic procedures are supported by a variety of other methods:

- Constant communication between the lab instructor and the child's regular teacher to secure information about weaknesses identified in the regular classroom.
- Constant monitoring of the child through periodic quizzes and review exams.
- Contact with parents for insight into the needs of the child.
- Constant observation.
- Securing the confidence of the child to enable him to express his own needs.

Instructors should work with a small number of children and should be assigned complete responsibility for the children's progress in both reading and math. The small groups, along with careful record keeping, enable the instructor to develop and maintain an in-depth understanding of each child's needs.

The assignment of materials that are relevant to the exact needs of the child is possible once specific weaknesses have been identified. Materials must be carefully assigned because the constant student success that Catch-Up tries to foster can be attained only if materials are geared to the individual child. Reading assignments are primarily keyed by the criterion-referenced materials, since they provide an index of some materials designed to attack specific deficiencies.

A list of additional relevant materials developed by instructors at the original site augments those indexed by Random House and is maintained by a card index system keyed to the criterion materials.

The positive, success-oriented techniques come into play after needs have been identified and relevant materials located. Careful planning assures that the child can achieve constant success and therefore a positive view of his own capabilities. The following procedures ensure the positive atmosphere of the program:

- The materials should be relevant, success oriented, and of a large variety.
- The difficulty and number of assignments should be based on the current learning style of the child and should be increased only as indicated by the child's progress.
- Constant monitoring and observation of the child are required for noting any significant changes in the child's needs.
- Clear and complete written progress reports should be maintained for constant reference.
- The teacher should be very positive and should refrain from negative comments such as "You're wrong."
- The teacher should "join in" with the students on games or discussions and thereby show the student that he can beat even the teacher on some things.
- The instructor should know the student so well that he can help in noninstructional areas, which also influence the child's self-image. The PIP specifically mentions becoming a friend and helping with regular classwork, resolving conflicts, and attempting to convey the child's needs to the parent.
- All activities should be carried out in a lab environment that is designed to add to the positive atmosphere. The lab should be bright and colorful with many displays, including the children's work and ethnic themes.
- The children should be encouraged to become partners in the education endeavor by being allowed to understand why specific assignments are made each day.
- The PIP also states that the instructor should be free to choose any instructional "approach," but this should probably read instructional "materials," since materials can be used in any way the instructor deems practical. Experimentation should be encouraged, and each teacher should have his own funds for purchasing materials he likes.

All of this instructional activity is provided with the understanding that, unless the child is interested and enthusiastic about what he is doing, he will not be motivated to learn. The operating procedures and the materials selected should be planned with this in mind. The main procedures to be followed are:

- A variety of materials should be available and utilized. Some materials intrigue some children, but not all. The instructor should use the materials that are right for the given child.
- Teaching machines should be used, since some children are intrigued with the novelty and as a result begin to like their work.
- Games should be available because they add to the excitement of learning.
- Effort should be made to avoid any overlap with materials used in the regular classroom. This maintains the freshness of the lab and prevents the child from feeling that the lab is just an extension of the classroom.
- Some free choice of activities and materials should be offered so that the child may express his preferences. Given the careful planning and specificity of assignments, free choice of materials is possible only when assignments have been completed or when a choice exists between relevant materials. Some limited time may also be set aside for free reading or extra time at the machines if it does not interfere with the schedules of others.
- Children should be allowed some privacy at times--for example in portable carrels or a somewhat isolated, quiet, reading-learning center.

Only by employing the careful procedures outlined above and carrying them out in a positive atmosphere could the original lab attain the gains they did.

4.3.2.2 Conquest Instructional Program

Project Conquest is a supplemental reading program designed from a comprehensive clinical point of view and aimed at bringing a remediable student up to grade level in reading. The design of Conquest is based on the following premises:

- Many students are not progressing in reading at the rate possible or to their potential.

- A pull-out program with a diagnostic-prescriptive approach can assess the needs of the student and then "cure" them.

- A reduced student-teacher ratio provides an opportunity for individual attention and an individualized program of instruction.

The instructional treatment for students consists of three phases: diagnosis, prescription, and remediation. Students are referred to the labs for initial diagnostic screening that incorporates a range of diagnostic reading tests, as well as auditory and visual screening to assess reading handicaps and needs.* A profile is then developed for each student; it should contain a graphic and/or numerical representation of the student's needs. Those who are one to two years behind grade level and show potential for remediation (upgrading in reading to grade level or as close as possible) are selected for a year of treatment in Conquest.

The second phase--prescription--then begins. A procedure for smooth transition from diagnosis to remediation was not satisfactorily provided by the PIP; however, it states that the child's program should be tailored to the child's needs. The PIP allows clinicians to use their own professional ingenuity to develop an individual prescription for each child's instructional program.

After prescription, remediation begins. Clinicians work with six students, 50 minutes a day, for four or five days a week. Each session, the Conquest student should experience three or four activities in the following areas: programmed reading, comprehension, phonics/vocabulary/sight words, and oral and/or recreational reading. Instruction in at least one of the areas should be assisted through the use of a teaching machine. The clinician is directed to personalize attention, to motivate, and to provide instructional situations that will ensure some success for the student. Extensive planning and record keeping are expected of the clinician so as to maximize instruction and minimize waste of time. Clinicians must keep folders, daily record sheets,

---

*Diagnostic screening is considered a continuous process. As students gain skills in subareas, they should be rediagnosed for other deficiencies and then remediated in those areas.

commercial record sheets, and detailed lesson plans for all students. Game days held for the students at the end of each week as a reward for working hard should be used to reinforce skills taught in a different manner during the week. In general, the clinician's role during an instructional period should be as follows:

- The clinician should be able to monitor the work of all six students, provide individual attention as needed, and still work with small groups when appropriate.
- Whether in a group, or on a one-to-one basis, the clinician should observe the behavior of each child and teach to the child, rather than to the group.

Students' behavior during an instructional period should reflect the planning and teaching style of the teacher. Students should be diligent workers, as evidenced by the time spent working on their tasks. They should enter the classroom knowing what is expected of them in terms of work. They should often work independently in carrels, after having built up self-confidence and the foundation for independent study. The student should usually begin the session with programmed reading--keeping detailed records of his own progress, and reading his daily record sheet for assignments in other subject areas. A student should move from one activity to another with a minimum of wasted time.

The instructional program, even within the previously described guidelines, may vary somewhat from lab to lab; however, the approach should be consistent among all instructional staff. Students in all labs should undergo diagnosis followed by an individualized instructional program that includes a variety of activities, a variety of teaching media, and careful record keeping.

4.3.2.3 High Intensity Tutoring (HIT) Instructional Program

High Intensity Tutoring (HIT) is designed to raise achievement levels in reading and math skills for middle school age students who are achieving one to five years below grade level. Tutees are sixth graders and some seventh graders; tutors include eighth graders and some seventh graders. Tutees are chosen from candidates who perform farthest below grade level and are selected on the basis of spring test scores and teacher recommendations.

A school participating in the HIT program should have one reading center and one math center, each staffed with one teacher and two aides who act primarily as instructional facilitators and bookkeepers. Tutors for the project should be volunteer students, who themselves perform at least one year below grade level.

Tutoring sessions are 30 minutes and are divided into 10-minute drill periods and 20-minute workbook sessions. Each tutee attends one tutoring session four times a week. Ideally, there should be twice as many tutors as tutees so that enough tutors are available for each tutoring session.

Each tutor sits at a desk close to his tutee so that both can see the materials being used. Furniture should be placed so that teachers and aides can circulate freely among the pairs of students.

Folders of materials should be available to each tutoring pair at the beginning of each tutoring session. The tutor should always be ready to prompt or encourage his tutee, but should never lecture or over-explain. He should use brief questions or corrections instead. The tutor should keep track of correct and incorrect answers by means of a tally sheet, using a slash for a correct answer, a zero for an incorrect answer, and an X when he has encouraged or complimented the tutee. Correct responses represent earned points to be entered into individual bankbooks by the staff daily and redeemed on "paydays" for rewards.

Some tutoring sessions should be designated as game days to provide a change of pace for both tutees and tutors.

Tutees and tutors should be motivated in their work by rewards, as well as by encouragement and attention to their progress and needs in the classroom. The rewards are candy for the tutees and field trips for the tutors. Paydays, already mentioned should be held once or twice a month for tutees, so that they may exchange points earned in class for the rewards they want. Different rewards have different values. During paydays, tutees may spend whatever points they choose, and a new balance is entered into their bankbooks.

Adult staff should circulate among the tutoring pairs, being constantly aware of each tutee's progress and problems and of the materials being covered. Staff should always be ready to reinforce a tutor's approach or give help when it is needed. Adult staff must also be aware of the degree of rapport between tutors and tutees and know when a tutor is not doing well with a tutee. Adult staff should take turns tutoring if too few tutors are present.

Staff must record daily each tutee's progress by means of an individual chart. That is, if the number of correct answers for a tutee falls below 90% for three days in a row, the difficulty of the materials is decreased and the student goes to a less difficult section. If the number of correct answers rises above the 94% mark for three consecutive days, the difficulty of the material is increased. At the end of a section of instruction, a test covering that level of material is given. If the student scores 85% or better, he is allowed to begin a new level; if not, the tutee must review the material again.

The materials used for the 10-minute drill are Hegge, Kirk and Kirk word lists and math flash cards. Sullivan reading and math workbooks are used during the last 20 minutes of the tutoring session. Teacher-made materials as well as games should be used as alternatives for the tutees when they need a change in routine.

The HIT reading component emphasizes basic phonics; reading comprehension is not part of the program. The following words are samples from Drill 6 of Hegge, Kirk and Kirk, in which a particular sound for the letter u is being learned: hut, run, pup, mug, rub, mud, and hum. The words are first shown spaced well apart from each other to facilitate reading of them. Gradually the words are placed closer to each other. The print is clear and easy to read. Later, in Drill 7 (a review), words containing the u sound will be interspersed with other sounds in words like cat and sip. To pass the review, the tutee must be able to read this combination of sounds with ease.

A Sullivan reading workbook might consist of sections from which the tutee reads sentences aloud and answers questions as to the content of the material. An example of a sentence for which a tutee might fill in letters or words is as follows: "A bellboy works in a hotel. This b_llb_y is taking the woman's l_ggage to her room." In the next series of pictures, different letters or words will be missing. Discriminatory responses are also used. In one picture a man dressed in an ordinary business suit and another man dressed in a cowboy suit are shown, and the sentence underneath asks whether the man on the left or on the right is dressed as a cowboy.

The math flash cards are designed with the problem and answer on one side and the problem without the answer on the other side. During the first 10 minutes of the session, the tutee uses the flash cards to learn number facts. The tutee may first go through the flash cards reading aloud from the side that includes both problem and answer. The tutor then holds up the flash cards one at a time showing the side with the problem only, and the tutee furnishes the answer. In the remaining 20 minutes, the tutee works out math problems by hand in the Sullivan math workbook. The tutor uses a paper slide to cover and then reveal answers as the tutee goes from one problem to another.

As the tutee learns the simple skills in reading and math, problems and sentence structure become more difficult. Increases in level of difficulty are accomplished by skipping pages in the books, and decreases are accomplished by reviewing pages. The HIT center teacher should review the performance of each tutee daily and make all decisions regarding instruction.

The atmosphere of the HIT center should be enthusiastic and task persistent. There should be a busy murmur of voices as tutees say their words and number facts aloud. They should give responses at a consistent pace--fast enough so that interest is maintained, but slow enough so that tutors can tally easily after each response. The tutee learns by actively practicing the skills he is learning, not by listening to explanations from his tutor or from adults.

4.3.2.4 Intensive Reading Instructional Teams (IRIT) Instructional Program

The Intensive Reading Instructional Teams (IRIT) project is designed to raise the achievement level of pupils who are deficient in the basic skills of language and reading. IRIT also attempts to improve the self-image of the students and to develop motivated self-directed learners. IRIT teachers guide students toward developing an appreciation for and pleasure in reading.

The IRIT instructional year is divided into three 11-week cycles. Students are enrolled for one of these three cycles, and 45 students are enrolled for each cycle. Instruction is provided each morning, for approximately three hours.

The IRIT program has three teachers, each with his own classroom and each instructing in a different area of reading. The 45 students are divided into three heterogeneous groups of 15 students each. After each 50-minute period, these groups move from one of the teachers to another. Thus each teacher sees all 45 students during the morning.

The areas of reading that are taught by the three IRIT teachers are decoding, vocabulary-comprehension, and individualized reading. Each teacher's room should be different in materials used and particular reading skills taught. Since students are exposed to an intensive reading program for the entire morning, they should be exposed to variability among the three classrooms. There is of course some overlapping in skills, but the IRIT teachers should attempt to identify these skills and assign each to one of the three classrooms.

Although each classroom is unique in materials used and skills taught, IRIT teachers are a team and a continuity should exist in each child's instructional program. To accomplish this, IRIT teachers should meet regularly to discuss the progress of each student and to inform each other of specific needs of individual students.

The key to IRIT is individualization. To individualize is no easy task, since teachers have 45 students per cycle for whom they must prescribe instruction. Moreover, every 11 weeks a new group of students is enrolled in the program. Their needs must be diagnosed, and the best way of helping them must be identified.

Careful diagnosis of reading problems is the foundation of the IRIT instructional process. The first week of each cycle is devoted largely to testing so that team teachers will have accurate diagnoses on which to base their lesson planning. The diagnostic instruments recommended in the PIP are the Batel Phonics and the Craft Word Attack and Comprehension Tests. An additional test used is the Random House Criterion Reading. Each team teacher is expected to select and administer diagnostic tests appropriate to her area of specialization. Each may also use instruments built into the instructional materials her students are using. In still other cases, she may develop her own tests to meet specific student needs. Clearly, the IRIT teacher is expected to come to the project with a good grasp of the role of diagnostic testing in the teaching of reading.

The teacher must also be familiar with specified commercial materials. After the student's deficient skill areas have been identified, the teacher assigns those materials that will help the student improve the specific skills. IRIT teachers should not automatically assign programmed materials to their students, having them "run" through materials designated by the publishers. Rather, the teachers should use all materials as tools for individualization. An IRIT student might be assigned two or three different materials that instruct in the same skill, but all materials should be reviewed and identified by the teacher as helpful for that particular student.

As well as helping students in those areas in which they are deficient, IRIT teachers must be aware of the need to help students grow in those areas in which they are strong. The IRIT teacher might have several materials that the students enjoy and might assign these on a regular basis.

Because of the intense individualization of IRIT, each student must work independently on his assignments. When students enter a classroom, they should go immediately to their individual folders, which hold their assignments for that period. IRIT teachers have a full afternoon for planning, and need it, since they must put these daily assignments in 45 folders. Each student should receive two or three assignments for one period; however, the number may vary, depending on the materials. Some IRIT teachers feel it best to allow students to work for a longer period of time with certain materials.

Occasionally, students will work in small groups--for example, when playing a language skill game, or when the teacher has identified several students who could benefit from working with her in a small group. Usually, however, students work independently. The teacher should walk around the room monitoring students, stopping to help as needed. Students should feel free to approach her for help. Every minute is important in the IRIT program, and teachers should see that students do not simply sit, wasting time.

A feeling of excitement should permeate IRIT classrooms. The students know that their assignments represent individual attention. They also know that they are expected to be independent workers, but should always feel free to ask for help when needed.

During the first week of the cycle, students must be taught how to use the various teaching machines and materials available in the IRIT classrooms. They must also learn the classroom routine and the procedure for recording their own progress in their folders or on wall charts.

The IRIT morning should be busy for both teachers and students. The many materials and machines, the three unique classrooms, and three different teachers provide enough variety to motivate students to stay with the work for the entire three hours.

A look at the three IRIT classrooms reveals the following distinctions. The decoding classroom is perhaps the most oriented toward very specific skills. In this classroom, students should be helped with any difficulties in phonology and should do a great deal of drill work. Here it is important that the decoding teacher use a variety of materials in new and interesting ways to maintain student enthusiasm.

The emphasis in the vocabulary-comprehension classroom is on reading comprehension. Students should be assigned a variety of reading materials, often accompanied by work sheets or other assignments. The Random House Criterion Reading has been most beneficial in identifying skills related to this room. The teacher must guard against simply scheduling students for the use of various programmed materials; she must use the materials in a prescriptive manner. Because activity in this room deals with skills that overlap activity in the decoding and individualized reading rooms, the team must decide where these skills should be assigned for treatment.

Of the three classrooms, the individualized reading room allows the most freedom. In this room, the teacher should attempt to motivate and guide students toward an interest in reading. Students should be given freedom to choose books of interest, either from a variety of books at a specific reading level or from a variety of books at all reading levels. A visit to this room should show students reading a variety of books, some of which have accompanying tapes. The teacher should have conferences with students (preferably with one student at a time) after books are read. Students should mark on a wall chart or in their personal folders any work they have completed and indicate if they are ready for a conference with the teacher.

4.3.2.5 Programed Tutorial Reading (PTR) Instructional Program

The objective of Programed Tutorial Reading (PTR) is to improve the reading ability and self-confidence of underachieving first grade students. Past evaluation of PTR programs indicates that, when properly implemented, the program has been an effective supplement to conventional classroom teaching. Indications are that it has been most effective with students who fall in the bottom quartile on national test score distributions in reading.

The instructional setting consists of a tutoring station that allows side-by-side seating for the tutor and the tutee. The tutoring location may be in a corner of the regular classroom, in a vacant classroom, or in any available school space that is free of disturbances.

Each student is tutored for 15 minutes each day by the same full-time paraprofessional tutor for the duration of the school year. During the 15-minute session, the tutor should adhere to tightly designed tutoring programs that carefully delineate and control instructional patterns used and should limit all decisions about a student's performance to judging the correctness of each response. Throughout the 15-minute session, the tutor records the student's failure for each reading item within a lesson on a record sheet so as to determine which items are to be presented again on succeeding attempts. In the final analysis, everything the tutor does and says is determined by what the record sheet indicates the student knows or does not know.

At present, materials used in PTR are available from six publishing companies in the form of tutoring kits. The materials from each of the publishers have been designed as supplements to pre-primer and basal readers used in regular first grade classrooms. Each tutoring kit consists of the basal reader used in the regular classroom, a comprehension and word analysis book, word list cards, record sheets, and a tutor's guide. The tutor's guide specifies teaching procedures in detail. It also contains a master list that specifies the order in which tutoring lessons are to be presented. A tutoring kit is needed for each tutor.

The Alphabet Skills booklet, published by Indiana University, is recommended as a prelude to the tutorial kit for children with no previous reading experience in kindergarten.

PTR instruction is both methodical and repetitive. It is dictated by 11 different tutoring programs called Item Programs (e.g., sight reading, reading, question, completion, and story), which were developed to supplement reading skills such as word analysis, comprehension, oral reading, and sentence construction. Each program specifies in detail what to teach and how to teach. For instance, all reading items taught in a given program are presented in the same format so that they can be taught with the same procedure.

Although PTR is programmed tutoring, the teaching strategy employed should be quite different from conventional programmed tutoring. Rather than seeking errorless or nearly errorless learning by providing initial cues followed by a fading of cues, the PTR tutor should practice minimal cuing at first, followed by increased prompting until the tutee determines the correct response. For instance, on each lesson the student should be presented all of the items within a lesson. After the first attempt (known as a run) at all items in the lesson, the sequence of successive steps is determined by the child's pattern of success or failure on items in the first run. If errors occur, additional runs should be presented, giving only items missed on preceding runs, until the student completes all items satisfactorily. Then all of the items in the lesson should be presented again and the process repeated until the student completes all of the items correctly in succession, or until ten runs have been attempted. The tutor then goes on to the next lesson.

The order in which the lessons are presented is dictated by the Master List. The general pattern of lesson presentation is cyclical: Several consecutive reading lessons are followed by a few comprehension or word analysis lessons. The lessons increase progressively in vocabulary, in length, and in task complexity, but each lesson builds upon skills acquired in previous ones.

An important component of the instructional process is frequent and immediate feedback. Each tutoring kit instructs the tutor to "Reinforce and go to Step __" following a correct response. Only positive reinforcement, which includes occasional use of the student's name, is used. Only the lack of positive reinforcement should be indicative of an incorrect response.

In summary, there are eight general aspects of programmed tutoring, as presented in the Tutor's Guide, Ginn 360 (one of the six tutorial kits available). Annotated descriptions of the general aspects are given below.

- <u>Programmed tutoring requires active learning</u>--The student in PTR learns by actively reading and reacting to what he reads, not by passively listening or being told. He reads and follows printed instructions, and he reads and chooses among alternative answers. Through the entire tutoring session, he should be actively learning as he practices the various reading skills.

- <u>Programmed tutoring is individualized teaching</u>--The rules that the tutor follows are the same for every student, but these rules allow the tutor to treat each student differently, depending on his individual ability as reflected in his moment-to-moment successes or failures in reading and in comprehension. As a result, programmed tutoring allows each student to progress at his own rate.

- <u>Programmed tutoring requires learning by discovery</u>--Each program is designed to help the student discover the answers to reading problems or questions by himself, rather than having the tutor tell him the answers. Each program begins with a test that presents the student with a reading problem or task in its most difficult form. If he cannot solve the problem in this form, it should be progressively simplified by changing its form or by providing more information, hints, or additional context. However, these changes and prompts should never provide a complete solution to the problem, so that in doing what a test step requires there is always some element of discovery.

- Programmed tutoring emphasizes success--Throughout all of the programs, a student's successes should be emphasized by praise and encouragement. His failures should be ignored in the sense that the tutor does not call attention to them (other than recording them on the record sheet). The student is simply taken to the next procedure in the program, usually one designed to teach the correct responses.

- Programmed tutoring provides the child with clear evidence of progress--Each student's progress is tied to his own successes, which should be clearly emphasized for him. Each student competes only with himself so that those who progress slowly should not be discouraged by comparison with others.

- Programmed tutoring is systematic teaching--The overall objectives are to teach sight-reading, comprehension, and word analysis. Each of these objectives is systematically represented in a revolving sequence of lessons. Each lesson builds upon previous ones, and each lesson must be mastered before going on to the next.

- Programmed tutoring is efficient teaching--Teaching time is concentrated where it is needed. Each student should progress quickly through material that is difficult so that a minimum of time is wasted in "teaching" what he already knows.

- Programmed tutoring is human teaching--Ideally, the good teacher is patient, sensitive to the student's need for success, tolerant of failure, and painstaking in matching her teaching procedures to the requirements of the individual student. In PTR, these virtues have been programmed.

4.3.2.6 R-3 Instructional Program

The goal of Project R-3 is to improve junior high school students' reading and math skills and, by providing them with success experiences, to improve students' self-image. This will help them to succeed in school and in the world of work. Toward this goal, the program employs individualized instruction in math, reading, and social studies, in combination with a laboratory approach using learning centers. What distinguishes R-3 from other individualized instruction programs is its emphasis on motivational field trips and gaming/simulation activities designed to demonstrate the applicability of classroom learning to problems faced in the world outside of school.

For three 45-minute periods each day, all students at one junior high school grade level receive instruction in math, reading, and social studies classes in which each student is on an individual progress program. Diagnostic tests are administered in the three subjects throughout the school year to determine each student's areas of strength and weakness. Based on the diagnostic tests, a program of instruction that will meet individual needs is prescribed for each student.

Learning contracts are the substance of the individualized instruction program. Contracts that emphasize specific skills and concepts are negotiated with each student. In conference with the teacher, the student agrees to complete a certain amount of work over a designated period of time (usually one week). Although contracts are used in all three subject areas, they differ somewhat in format and use from one subject to another.

Math contracts for each content area (e.g., multiplication, fractions, percents) consist of several subdivisions of instruction and exercises in graduating degrees of difficulty. When the student has completed a section of the contract, he is tested on the work covered. If there are skills he has not yet mastered, they should be reviewed and additional exercises assigned. When he is able to complete the final test on a contract successfully, he moves to the next step of his individual program. A variety of reinforcing activities should be used in conjunction with the contracts. For example, programmed instruction kits are available for supplementary work, and each classroom should have a supply of games that require the use of the particular skills the student is acquiring.

Reading contracts are developed around a central theme (e.g., science fiction) and encompass activities for all ability levels, from which assignments appropriate for each student should be made. Each activity carries a specific number of points, and the number of points earned determines the student's grade. When a student successfully completes his individual assignment, he can choose to do additional activities on the contract to increase the number of points he earns and thereby improve his grade. A reading contract might include several activities on different ability levels under each of the following categories: exercises in punctuation, spelling, vocabulary, dictionary skills; readings from one or several literature series with questions to answer; original compositions; games that require the use of specific skills; and crossword puzzles. The materials that students need to complete the various parts of the contract should be set out around the room at learning centers where students can work together or individually, as the activity dictates.

Social studies contracts are similar in format to the reading contracts, and students carry out assignments for points (and grades) at learning centers. A contract might include tasks such as report writing on suggested topics, readings with corresponding questions, games such as matching people to events, an exercise related to the vocabulary in the reading, and perhaps a math activity related to the unit being studied.

Games and simulations are used principally in social studies classes, although they are also used in reading and math as motivating and reinforcing activities. They should also be used to illustrate the relevance of material learned to situations that are encountered outside of school. Since simulations are designed to integrate the three subject areas and to reinforce skills the students have learned, they make use of math and reading skills with a social studies theme. For example, the gaming/simulation activity called "Hurricane Warning Game" is used during a social studies unit on weather. It instructs students to make decisions about whether to secure their towns against Hurricane Eva, which is hovering offshore. Each student selects a specific town for which he is responsible. Knowing the cost of securing the town and the dollar amount of possible damages as well as the probability that the hurricane will strike his town, the student must make his decision based on the computation of probable savings of securing versus not securing the town. He then throws dice, the sum of which will determine whether the hurricane struck his town or not.

Simulations that incorporate difficult math concepts or vocabulary should be introduced in math and reading classes, where instruction can be given to prepare students for the activity.

Like the gaming/simulation activities, field trips are an important motivational force in the R-3 program. Trips should be arranged from time to time as funds permit to point up lessons learned in the classroom. For example, after map-reading and route-planning lessons, students may be taken on a bus trip they have planned themselves to places of interest in their city.

Extended field trips of two or three days, called intensive involvements, should be planned for fall or spring to a suitable site where students and project staff can spend one or two nights away from home and school. Intensive involvement is a series of learning experiences built around a particular theme and includes gaming/simulation activities. Preparatory activities should be conducted in advance of the study trip, and follow-up activities should build on the experiences of the students at the intensive involvement site. Such trips become the highlight of the year for both teachers and students, many of whom have not been away from home previously. The informal atmosphere of the out-of-classroom experience encourages closer relationships between teacher and student than can develop in the more formal classroom.

All classes should offer a variety of instructional approaches. Lessons are given individually, or in small groups, and sometimes to the entire class. While most of the contract assignments are carried out individually, some tasks require students to work together in small groups. To avoid the possibility of students becoming bored with working independently, teachers should periodically devote a week or so to activities that involve the classroom group as a whole. Because the individualized instructional program requires teachers and aides to spend so much time with individuals, students should sometimes be divided into small groups with student leaders who answer the questions of their group members; this approach allows the teacher and the aide to devote their time to students who have more serious problems.

4.3.3 Conclusion

The completed program descriptions, which reflected the modifications resulting from our own internalization of the PIPs, as well as the revisions and clarifications agreed upon at the Washington conference, gave us an overview of what the tryout projects should be like during the second-year site visits; however, areas of instruction for some PIPs were still unclear. The PIPs lacked detail about some of the procedures, especially how the materials were to be used and assigned and how an individualized curriculum would unfold in a well-implemented classroom.

4.4 Site Visit 1

4.4.1 Data Collection

As stated in Section 4 1, for our first site visit of the second year, we selected a sample of teachers to interview and classrooms to observe so that we could pursue implementation of the instructional programs in depth, rather than repeat the first year's evaluation of the general level of implementation. Our evaluation plan was based on the assumption that the major program characteristics observed in the first year of implementation at each project would remain much the same in the second year and that little new information would be gained from a second-year analysis in which data were accumulated on every teacher at every project. It seemed likely that more information would be revealed about the effects of instruction on students' success or failure if we looked more closely at only a few classrooms or treatment groups, particularly those that fell at the extremes of implementation. Therefore, we chose those treatment groups that had indications, on the basis of the first year's observations, of being either rather well or rather poorly implemented.* Since the first year's observations had focused on how closely teachers followed the PIP specifications for setting up and managing the classroom, how they interacted with students, and the clarity of their presentations of lessons, our implementation judgments were made after a consideration of those areas of the program. Table 4-2 shows the number of instructional staff observed and interviewed.

Many features of the instructional program were not well described in the PIPs, but were nevertheless essential for a site to reproduce in order to call its program implemented. The specifications in the PIPs were concerned mainly with space, furniture, instructional equipment and materials, and the adult-student ratio. Our observation procedures had to be sensitive to the classroom processes that were likely to influence achievement test scores. Otherwise we would have no documentation to justify the expectation that the MAT results would be relevant to PIP impacts.

The program descriptions presented in Section 4.3.2 were used to guide the development of PIP-specific observation instruments for the first site visit. Although the instruments were PIP-specific, each followed the same general plan, which allowed us to document the instructional treatment experienced by students and the implementation of PIP specifications for both classroom management and curriculum resources. After pretesting and revision, the final observation instruments included

---

*In R-3 sites where many of the teachers were new to the project in its second year of implementation, such a judgment could not be made. Therefore, the sample for these sites was chosen on the basis of fitting some teachers from each subject area into a workable interview and observation schedule for the site visitor.

Table 4-2

OBSERVATION AND INTERVIEW SAMPLE

PIP	Project	Number of Teachers in Sample
Catch-Up	Bloomington	4
	Brookport	2
	Galax	2
	Providence Forge	2
	Wayne City	2
Conquest	Benton Harbor	3
	Cleveland	3
	Gloversville	4
HIT	Lexington	3
	Olean	4
IRIT	Bloomington	3
	Oklahoma City	3
	Schenectady	3
PTR	Canton	6
	Dallas	6
R-3	Charlotte	7
	Lake Village	3
	Lorain	5
	Schenectady	5
Total sample		70

means for documenting the following indicators of implementation of instructional techniques:

- Classroom features and resources.
- Classroom management before and during the instructional period.
- Student grouping arrangements.
- Individualization of materials and subject matter.
- Summary of teaching techniques of the instructional staff.

- Summary of student behaviors and responses in the instructional setting.

Most of the sections of the observation instrument were designed for straightforward recording of occurrences; however, some sections necessarily required more judgment by the observers.

One of the areas in the PIPs that was generally left unspecified, or was at best unclear, was that of how the curriculum materials should be used in an individualized instructional program. The section of our observation instruments called "Individualization of Materials and Subject Matter" required the observer to record the materials that selected students were working on during the class period. Before we observed in classrooms, we asked the teachers to consult their records for four or five students and to tell us what materials and lessons they would be assigned during the class period in which we would be observing. With this information, we could verify whether students were working on different materials and levels (as prescribed by an individualized program) and whether the selected students were in fact using the materials assigned to them. We could also determine whether students were using the materials appropriately, that is, whether they were actually engaged with and working on the materials. From these observations, we could judge whether or not learning was taking place in the classroom.

In our interviews with teachers, which followed the observation period, we identified changes that had occurred since the first year relative to roles, training, and the use of materials. We also questioned them about their interpretation and implementation of PIP concepts such as "individualization," "core materials," and "diagnostic-prescriptive procedures." We asked them to describe a typical instructional period, including how they made assignments, what materials they used and why, and details of their record-keeping systems.

In our view, we could not make a plausible argument that the PIP projects would influence MAT scores unless we investigated the curriculum that the teachers were using. For this reason we completed, on the first site visit, a "Dictionary of Core Materials" with each teacher we interviewed and observed.* The dictionary listed and described the materials specified in the PIP, as well as materials that had been observed

*In R-3 sites, the dictionary was completed on math materials only because the PIP specified a more manageable number of math materials than reading materials.

in classrooms during the first year of the field test. The site visitor and the teacher located materials in the classroom, and the site visitor checked them off in the dictionary. Teachers were asked to designate materials they considered to be "core" materials which, for this purpose, we defined as materials used with 50% of the students or used 50% of the time. In addition, we asked teachers how they divided the materials into teaching segments (i.e., chapters, sections of chapters, particular pages) and recorded that information in the dictionary. The completion of a "Dictionary of Core Materials" for each classroom enabled work to begin on a curriculum analysis that would tell us whether the PIP projects used materials whose content was covered in the MAT battery.

4.4.2 Review of Data from Site Visit 1

After the site visits, we attempted to use the data we had collected to classify projects relative to their performance on the instructional component of the PIPs. We had, however, made two errors that prevented us from satisfactorily doing this. The first error was that we had thought it possible to assess implementation globally, at the project level. The PIPs had been designed to create projects, and this seemed the appropriate level of abstraction at which to evaluate implementation. Unfortunately, at that level of abstraction, the degree of implementation was not directly observable. The site visitors were unable to verify directly that the project was implemented; they could verify only that some of the project staff were behaving as specified in some respects, but not in others.

Since what was observable were project staff interactions, the tendency was to refer to how well individuals were doing. At this level of detail, site visitor reports were objective, in that they were based on direct observations. At the project level, their reports were clearly subjective, in that they had to "sum up" their judgments about the staff to reach a conclusion.

Our second error was that of not putting the descriptions of the instructional component of the PIP on our observation instruments in the order in which instruction usually occurred. We found it difficult to determine the success of a teacher on activities that we did not anticipate would occur. It was also difficult to distinguish events that were not specified in the PIP from ambiguities, the teacher's resolution of which interested us.

The next section describes the steps we took to overcome these weaknesses in our original procedures.

4.5 Identification of Gaps and Ambiguities in the PIPs

Since a major objective of this year's field-test evaluation was to find those aspects of each PIP program that contributed to improvements in student test scores, we felt a need, based on the interviews and observations of the first site visit, to reorganize our descriptions of the instructional components for each PIP program. The organization used on the original site visits made it difficult to verify directly that the specified techniques were being used. Reorganization would give the details of the instructional procedures in the order in which they were used in the classroom. Such an exercise would force us to think through areas that were still vague and would perhaps give us a better grasp of the programs for our final visit to the project sites.

We also hoped that with the new organization we could more easily discriminate variations in the learning/teaching process across sites within PIPs. These variations were expected to be explanatory factors for the effects of the PIP projects on student achievement, as well as suggestive of revisions for the PIPs, which were being rewritten for dissemination.

As an outline for reorganizing the descriptions, we used a typical teaching plan* that broke down the instructional functions into the following steps:

- Diagnosis of each student's needs
- Prescription of materials designed to meet diagnosed needs.
- Presentation of the lesson/skill
- Guided practice; independent study
- Assessment of progress (monitoring, testing)
- Reinforcement activities
- Motivating techniques.

As we organized the descriptions into these steps, the programs took on the essential structure and order that had been missing previously, yet there were still areas about which we were unsure. What did become quite clear, however, was where the PIPs were vague or allowed

---

*A teaching plan used by the Cleveland Conquest instructional staff was adapted for this purpose.

the teachers a great deal of freedom, and what areas of instruction the PIPs omitted entirely. Now we could designate those aspects of the instructional programs that were ambiguous, free to vary, or not specified in each PIP. For each step of the teaching plan, we developed a matrix for each PIP on which we listed the revised descriptions. Discrepancies between our descriptions and the PIP guidelines were identified and explained under the column headings, "not specified," "ambiguous," or "free to vary," as appropriate. The matrix developed for Project Conquest is shown in Table 4-3, which illustrates the method used to classify descrepancies for each PIP.

The completed matrices served several purposes. They clarified for us the reasons why we, and the tryout project staffs, had felt secure with the management aspects of the projects, but had been baffled by the instructional programs. We had attributed our confusion to the organization of the PIPs, since directions were scattered throughout the components in sometimes unexpected places and in differing amounts of detail. Now we were able to show graphically the gaps and ambiguities in the PIPs that had resulted in our less-than-complete understanding of the programs. The matrices easily allowed us to identify weak areas in the PIPs and to generate suggestions for revisions that would help clarify the packaged programs for future users. In addition, they provided the basis for our final visit's interview questions and guided the development of our observation instruments for the second site visits.

Examples of gaps and ambiguities for each PIP are described below to give the reader a better understanding of the areas that we felt were of special interest for assessing the degree of implementation.

The Conquest PIP specified several activities that had to occur in the labs so that each student could participate in three or four activities daily. As shown in Table 4-3, the PIP did not present the details of classroom management that would help teachers plan the class period to incorporate the specified activities, as well as to give individual attention to each child, handle problems that might arise, and keep the detailed records that the program demanded. Since the implementation of Conquest's individualized instruction program depended on efficient classroom management, one of the speical foci of our second site visit interviews and observations was on teachers' orgainzational and planning skills.

The Catch-Up PIP recommended the use of a wide variety of materials along with the Random House Criterion-Referenced System for diagnosing needs and prescribing related materials. The use of the criterion-referenced materials, however, was not explained in the PIP, and the

Table 4-3

ASPECTS OF THE [illegible] INSTRUCTIONAL PROGRAM: NOT SPECIFIED, AMBIGUOUS, FREE TO VARY

Program Aspect	EIP-Specified	Omitted/Not Specified	Ambiguous/Contradictory	Free to Vary
Diagnosis	[illegible] step diagnostic process (modified at Washington [illegible], feedback conference).	Appropriate substitutions for tests no longer in print.	Errors and misprints in some tests.	
	Five-week diagnostic testing period	How to schedule testing to minimize frustrations for students, clinicians, and regular classroom teachers.		
	Diagnosis, a continuing process throughout the year as needed.	Interpretation of test scores.		
		When clinician should further diagnose the student.		
		How test score information interfaces with instructional program.		
		How to begin setting objectives for students; procedures or concrete suggestions.		
[illegible]	Students to be started 1 year below tested grade level. Difficulty gradually increased.	How to interpolate 1 year below tested grade level in specific skill areas.		
	Student profile sheet to be developed [illegible] as a basis for individualizing student's program.	How to individualize [illegible] student's program.		
	Treatment for grades 1-3 [illegible] should [illegible], basal text, programmed reading, oral reading, teaching machines, games.			How much of the subject area to include in the daily prescription for each child (as long as the subject areas are covered daily)--teacher's decision based on child's needs.
	Treatment for grades 4-6 [illegible] should include programmed reading, comprehension, vocabulary, sight words, teaching machines, recreational reading, oral reading.	How to match materials with students.		Materials to use with particular student to teach a particular skill.
[illegible]	Treatment time		Time only specified as: [illegible] 30 min, 4.5 sessions per week; 2 sessions per week.	
	Student-teacher ratio: 1 clinician to 8 students per period; 1 supervising clinician to 4 students (also substitute for absent clinicians).		Role of supervising clinician unclear.	
	Develop and implement lesson plans: determine student needs, decide how to motivate, match student with appropriate materials, give student [illegible] confidence.	How to develop lesson plans for teaching certain skills that correlate with objectives. How to present a specific skill to students. Discussion of differential treatment of younger as compared with older students.		

Program Aspect	PIP-Specified	Omitted/Not Specified	Ambiguous/Contradictory	Free to Vary
Guided practice	Guided practice with teacher; student is to receive some individual attention.	How to group students while teaching.		Whether teacher works with groups or wi[illegible] individuals--left t[illegible] [illegible] judgment.
	Use of carrels for independent work, for some activities when students are confident enough.	How to determine when students are ready. How to assess effective versus ineffective use of carrels.		How much time a particular student works in the carrel--teacher's decision.
	Use of teaching machines and devices; daily use advised.			What machines to use with each student.
	Student schedules and classroom management: How long students should spend on a subject area each day. Reading room --grades 1-3: phonics, 10 min; basal text, 10 min; programmed reading, 15 min; oral reading, games, teaching machines, 10 min. Clinics--grades 4-6: programmed reading, 10 min; comprehension, 10 min; vocabulary, 5 min; sight words, 5 min; teaching machines, 5 min; oral reading, 10 min.	How to manage the classroom within the guidelines. Amount of flexibility allowed among clinicians and among students. How to schedule students around school schedules.	Not clearly described.	Decisions regarding levels within an activity for each student--clinician's decision.
	Roles of clinician: monitor, tutor, diagnostician, motivator, observer, prescriber, organizer.	How to perform in the various roles.		
Progress assessment	Record keeping: folders, notebooks, daily record sheets.	How good records help teachers maintain awareness of student progress. Descriptions of such records. How records relate to objectives. Discussion of various techniques and importance of the record keeping.	Records descriptions vague. Notebook and folder contents ambiguous. Some commercial examples were in PIP.	
	Symptoms.	Meaning. Symptoms of success or problems.	Vague item called symptomatology found in Project Director's Directory.	
	Instructions to diagnose as necessary.	When to rediagnose students. How to diagnose after initial intensive diagnosis. How to use results of initial diagnosis throughout the year.		
	Release of the student from treatment: posttest; consultation between project director, clinician, and supervisor.	How to use posttest results. How and when student should be released from program.	Obscure. Mentioned in passing.	
Re[illegible]iforcement	Game day: use of games that require students to use skills tried during the week or at some previous time.	Other [illegible]cedures for reinforcement. How clinician provides intermittent review of skills previously taught.		
Motivating techniques	Game day: 1 day per week. Thursday afternoon and Friday morning suggested. Providing success experiences: Making assignments that students can do.	Methods of motivating students who do not respond. (Little information provided on how to motivate students at all.)		Additional means of motivating students when necessary--teacher's decision.
	Praising students for every little thing.		Found in an obscure place in the PIP.	
	Use of achievement awards: for graduation, extra reading, honorable mention, attendance.			Use of achievement awards as needed--teacher's decision.

Random House system was keyed to very few of the PIP-recommended materials. Presumably to help the projects with this problem, the PIP included information that had been reproduced from a set of index cards used at the originating site to identify those sections of certain materials that were relevant to specific skill deficiencies. Unfortunately, the index cards were also keyed to many materials that were not recommended in the PIP. Therefore, an analysis of each of the PIP's core materials would have to be conducted by project staff to determine how they related to diagnosed needs. For teachers inexperienced both with the criterion-referenced system and with the PIP-recommended materials, this would be a fairly complicated and very time-consuming task. Yet to implement the Catch-Up program as specified, the task would be essential. How the criterion-referenced, diagnostic-prescriptive procedures were handled was of special interest in the Catch-Up second site visit, since it was an area that would be considered in a judgment of teachers and of well and poorly implemented projects.

The HIT PIP had few ambiguities or omissions. The instructional program was specified in enough detail to enable project staff to meet requirements for tutor training and classroom management, and the PIP-recommended materials incorporated adequate explanations of how they were to be used. The principal omission was the lack of explicit specifications for the instructional pace that tutor-tutee pairs should maintain--an essential ingredient of the originating site's program. Because the pace at which lessons are conducted contributes to the degree of interest and enthusiasm with which students attack their work, the HIT observation instrument included an assessment of instructional pace in each center. The activities of the adults during the tutoring sessions were also recorded on the observation instrument, since the PIP specified that they should be circulating as monitors throughout the session, helping as needed, but not interrupting the tutor-tutee interactions.

The IRIT PIP recommended two basic diagnostic tests for the first and last weeks of the cycle and specified that teachers select and administer other appropriate diagnostic tests during the cycle. The PIP did not explain how to select tests appropriate to the subject matter and student levels or how often to administer the tests. Neither did it explain how instructional materials were to be used to teach the diagnosed skill deficiencies. Therefore, as focal points, the IRIT interviews incorporated items relating to teachers' choice and use of diagnostic tests, how they prescribed materials and activities for individual students, and how they maintained records of needs and progress for individual students.

Like HIT, the PIP for PTR was straightforward and, when used in combination with the specified self-explaining materials, gave enough details of instruction to enable the staff to set up and maintain the program with ease. For this reason, the PTR interviews and observations were focused on determining to what degree the specifications were followed, rather than on the resolutions of ambiguous statements or omissions.

The R-3 PIP specified that the individualized instruction program be structured around learning contracts and that they be individually negotiated with students. Individual negotiation encourages students to assume responsibility for their own work and gives them an understanding of the importance of setting realistic goals. Yet the PIP gave conflicting information about contract format and use: The PIP stated that R-3 Teachers were supposed to develop their own multilevel, multiactivity contracts, but, in fact, a series of math contracts, each devoted to a single area of the math program (e.g., fractions), which were developed at the originating site, were specified as part of the core materials. The PIP stated that each contract should incorporate a posttest, but the sample reading contract reproduced in the PIP had no posttest; the PIP stated that contracts should contain predetermined grading schedules offering the student a choice of grade to be achieved, but the core math contracts did not contain grading schedules. Interviews with the tryout project teachers included questions on the use or development of contracts in each subject area, and the observations recorded whether and how contracts were used in the classrooms.

Interview guides and observation instruments were revised for each PIP before the second site visit to focus more sharply on those factors that would differentiate teachers, not projects, on how well they implemented the PIP instructional component. When the site visits were completed, we used the data that were encoded in the field to assess the teachers' degree of implementation in terms of the PIP specifications, as clarified by the Washington conference. However, we did more than collect data that would allow us to judge implementation solely on clear PIP specifications; we also collected data that would allow us to judge whether projects resolved ambiguities in conformity with the project philosophy, as gleaned from the package and the conference.

4.6 Site Visit 2--Data Collection

The revised observation instruments were developed to record the functions of the teaching plan discussed in Section 4.5. We observed and recorded how the instructional process was handled in the classroom, including those areas that were ambiguous or not covered in the PIPs. We subsequently interviewed the sample teachers to learn how they handled these ambiguities. As on the first visit, each instrument was

PIP-specific, but all instruments were now organized into the following general categories:

- Classroom management and grouping arrangements
- Individualization of materials
- Classroom facilities and atmosphere
- Classroom features and resources
- Student behaviors during instructional period
- Teaching techniques of the instructional staff.

The observation form allowed us to register whether the PIP-specified structure was there and to document teaching practices that we judged "good implementation" or "poor implementation."

Examples of our collection procedures are shown in Exhibits 4-1 and 4-2, which are reproductions of records made in an R-3 classroom on second site visits. The students were working independently and in small groups at learning centers on contract activities that required reading, writing, research, and some math skills. Information was recorded about student interactions and working behaviors and about their activities and materials (Exhibit 4-1) as well as about how the teacher spent his time (Exhibit 4-2) during the observation period. The observation record shows that instruction in this classroom was being implemented as specified in our instructional program descriptions.

How teachers interpreted those parts of the instructional program that were ambiguously stated or missing in the PIP was the emphasis of the interviews, but we also questioned them thoroughly about how they managed each step of the program. We asked teachers generally how they planned an individualized program for each student and specifically how they determined, for example:

- On what skills to put each student to work.
- What materials each child should use.
- What instructional approach is best for each student.
- Whether to introduce new material to individual students or to small or large groups.
- Whether a student is learning or not and, if not, what to do about it.
- What motivates each student.

Exhibit 4-1

STUDENT BEHAVIORS DURING INSTRUCTIONAL PERIOD

(Most Students)	Time: Most	Some	None	As Specified	Not as Specified	Not Specified
1. Students feel free to approach adult for help	1	2	3	1	2	3
2. Students ask each other for help	1	2	3	1	2	3
3. Students wait for adults to offer help	1	2	3	1	2	3
4. Students initiate task-related questions	1	2	3	1	2	3
5. Students initiate nontask-related question	1	2	3	1	2	3
6. Students quiet and orderly	1	2	3	1	2	3
7. Students converse but do not disrupt class	1	2	3	1	2	3
8. Students disrupt entire class	1	2	3	1	2	3
9. Students appear interested in task	1	2	3	1	2	3
10. Students appear enthusiastic about task	1	2	3	1	2	3
11. Students appear bored with task	1	2	3	1	2	3
12. Students appear restless	1	2	3	1	2	3

13. Average number of activities per student during instructional period . . . 1 2 3 4+ N/A 5

14. Approximate minutes of each activity . . . (5) 1 (10) 2 (15) 3 (20+) 4 5

15. Materials and activities:

	Yes	No
a. Students work on individual assignments	1	2
b. Class works on a variety of materials	1	2
c. Students work on materials in small groups	1	2
d. Students work on materials in large groups	1	2

16. Activities Occurring:

	All	Some	None
a. Students are working on contracts	1	2	3
b. Students are working on games	1	2	3
c. Students are working on simulation	1	2	3

17. Comments: Yes 1 No 2

Students seem genuinely interested in their work -- they are working hard and are quiet despite task-related conversations among groups at the various learning centers.

Exhibit 4-2

TEACHER INSTRUCTIONAL TECHNIQUE

(Circle appropriate number)	Time: Most	Some	None	As Specified	Not as Specified	Not Specified
1. Teacher stays with students, helping as needed	1	2	3	1	2	3
2. Teacher circulates, helping students as needed	1	2	3	1	2	3
3. Teacher remains in one place and students go to adult	1	2	3	1	2	3
4. Teacher works with individual students	1	2	3	1	2	3
5. Teacher works with small groups	1	2	3	1	2	3
6. Teacher works with entire class	1	2	3	1	2	3
7. Teacher spends time with other adults (observer)	1	2	3	1	2	3
8. Teacher spends time at classroom management	1	2	3	1	2	3
9. Humor is evident between students and teacher	1	2	3	1	2	3
10. Aide spends time at classroom management	1	2	3	1	2	3
11. Aide helps individual students with assignment	1	2	3	1	2	3
12. Teacher corrects students' unacceptable behavior	1	2	3	1	2	3
13. Teacher gives positive feedback to students for their work	1	2	3	1	2	3
14. Teacher gives inappropriate feedback to students for their work	1	2	3	1	2	3
15. Teacher style	1	2	3	1	2	3
16. Teacher is a facilitator	1	2	3	1	2	3
17. Teacher is controlling	1	2	3	1	2	3
18. Teacher lacks control of class	1	2	3	1	2	3
19. Teacher participates in activities	1	2	3	1	2	3
20. Teacher disregards activity and/or children's needs	1	2	3	1	2	3

21. Teacher's Planning (Circle appropriate number)

Teacher's plan for the period is clearly evident in her interactions with the students 1

Teacher's planning is somewhat evident through her interactions with students 2

Teacher's lack of planning is evident through her interactions with students 3

22. Changing of Activities:

	Yes	No	N/A
Teacher displays flexibility in reassignment of work to individual students	1	2	3

Interviews and observation data were used to generate a classification scheme for describing those teacher implementations on which analyses of curriculum and instructional materials would be conducted, as described in the following section.

4.7 Assessment of Implementation and Teacher Responsiveness

Following the second site visit, we completed debriefing tasks in preparation for a final selection of treatment groups whose curricula and test scores would be analyzed in detail. The selection process required a judgment of whether the students whose classes we observed and whose teacher we interviewed were learning in the way the PIP intended, and whether the teacher's implementation of the instructional program was in accordance with the PIP's explicit specifications and with the program's philosophy of teaching and learning. These judgments were formed and justified by the information collected in our interviews and documented on our observation instruments. In making our judgments of implementation, we used that observation data for reference as we wrote a description of the instructional program of each teacher, emphasizing how the teacher handled the gaps and ambiguities in the PIP specifications and describing her/his implementation of each of the steps of the teaching plan. Then, from the descriptions of individual teacher implementation, we constructed a generalized description of the instructional program at each project, citing examples to support judgments and noting exceptions. Variations in implementation across projects within PIPs were designated and probable causes explained, including management, training, contextual resolutions, professional experience of the teachers, and the like. These summaries provided the basis for our recommendations for PIP revisions.

With the completed descriptions of individual teacher implementation and the generalized description of implementation at each project, we rated each teacher on the basis of overall performance on two dimensions:

- Implementation, as shown by:
  - Attention to the explicit specification in the PIP (evidenced by the degree of fidelity to the specified procedures and activities).
  - Understanding of the program philosophy (evidenced by the method of resolving the PIP's gaps and ambiguities).
- Responsiveness, as shown by:

- Awareness of each student's progress during the period.
- Awareness of each student's interests (evidenced by teacher having time for personal comment).
- Amount of time teacher was working with students (versus time spent on classroom management tasks).
- Amount of time students waited for individual attention. (Some sat with hand up for 5 minutes or more; others stood in line at teacher's desk for 10 minutes or more.)
- Quality of attention. (Some teachers listened carefully to students' explanations of problem; some jumped in and explained things the student already knew and left before the real problem was solved.)

Along both of these dimensions, teachers were rated "good," "so-so," or "poor." Below we present some specific examples, using the same areas of investigation referred to in Section 4.5 to state our rating procedure. To maintain the confidentiality of our data, we will speak in generalities. However, in fact, each teacher rating was considered in detail on a teacher-by-teacher basis.

4.7.1 Assessment of Implementation

How Catch-Up teachers handled the criterion-referenced system of diagnosis and prescription contributed to their final ratings. Some of the experienced teachers seemed to grasp the concept immediately and were able to systematize their procedures so they could easily designate the specific parts of a variety of materials that were relevant to diagnosed skill deficiencies. These teachers were rated "good." On the other hand, some of the less-experienced teachers never fully utilized the criterion-referenced system. They used their own informal assessment of students' strengths and weaknesses or depended on a few materials with which they felt comfortable to meet every student's needs. They were not able to coordinate the diagnostic and prescriptive steps in the lesson plan, so we gave them a "poor" rating.

The Conquest PIP failed to describe how teachers could organize their classes to incorporate the several activities and record-keeping responsibilities that were supposed to occur daily. Because Conquest's individualized program could not function well without careful planning, accurate and up-to-date records that relate objectives to tasks, and efficient management of a variety of simultaneous activities, all teachers were judged on their organizational skills. Good classroom management was necessary but not sufficient for judging a teacher "good," and other features of her implementation had to be considered in the final rating. However, poor classroom management necessarily implies a "poor" rating for implementation because a poorly organized Conquest program could not be well implemented.

The instructional pace of HIT tutoring sessions was implied but not described in the PIP. Since the rhythm of tutor-tutee interactions was a good indicator of whether learning was taking place, the implementation of HIT teachers was judged, among other things, on the ability to keep the tutoring sessions moving along at a pace that would encourage the active participation and enthusiasm that characterize High Intensity Tutoring. In a center where tutors knew their job and were conscientious and quick about cuing, reinforcing, catching errors, and tallying responses, and where the teacher circulated and helped but did not engage in lengthy explanations (which break the rhythm of the tutoring session), the teacher was given a "good" rating for his understanding of the program's intent. In a center where tutors and tutees seemed apathetic or engaged in off-task conversations, and where the teacher was attending to record-keeping tasks rather than circulating around the center, the teacher was rated "poor" on his program implementation.

Although the IRIT PIP did not give directions for using the diagnostic tests or the variety of materials it recommended, the diagnostic-prescriptive procedures and an efficient record-keeping system were essential to the IRIT individualized instruction program. IRIT teachers were interviewed in depth about their use of these procedures and were judged on their implementation of them. A teacher who was judged "poor" on this aspect of instruction did not use any diagnostic tests because he "felt" that his students needed help in all skills, and his prescriptions consisted of rotating the students through a set of selected materials. A teacher who was judged "good" on this IRIT feature used highly systematized diagnostic-prescriptive procedures that enabled him to specify what pages of a variety of instructional materials were related to skill deficiencies diagnosed in testing. His students were advised of the skills on which they needed to work.

Because the PTR instructional program was clearly specified by the PIP and the materials it used, it was possible to judge implementation without the confounding of ambiguities. Tutors were judged on how closely they followed the directions in the programmed materials during the observation period. Although the materials contained a carefully delineated test so that everything the tutor should do or say is explicit, including feedback, some tutors were judged "poor" because they deviated from the text, used informal teaching methods, and employed negative feedback. Their students were given confusing instructions and appeared to be intimidated by the tutors. Tutors judged to be "good," on the other hand, adhered to the programmed instructional methods, provided positive reinforcement, and demonstrated an interest in their students' work. These tutors seemed to have captured the spirit and intent of PTR.

The development and use of learning contracts in the R-3 program was one of the ambiguous areas on which R-3 teachers were judged. Some of the reading and social studies teachers, despite tight schedules that afforded little preparation time, designed contracts that included activities for a variety of achievement levels and a predetermined grading schedule, while other teachers neither designed nor used contracts. A few teachers attempted to negotiate contracts with students, but the negotiations were quickly dropped because students did not set realistic goals and the experience proved disappointing. The result was that those teachers who used contracts made some assignments of specific activities and then allowed students to do further work if they had time and wished to earn a higher grade. In effect, the students were learning to set their own goals and to take responsibility for their own work, so although the contracts were not negotiated, the solution was compatible with the intent of the program. Therefore, those teachers who designed and used contracts in this way were judged "good," while those teachers who used no contracts were judged "poor" on this instructional feature.

4.7.2 Assessment of Responsiveness

The responsiveness of the teacher was judged on her interactions with her students. For example, one teacher was judged unresponsive when one of her students sat idle for 15 minutes before being noticed. Another teacher asked a student to read aloud, then walked away before the student had finished; this teacher was judged unresponsive. In some classes, up to one-third of the children were observed to be waiting for new assignments.

One teacher judged "good" on responsiveness seemed to "have eyes in the back of her head." She gave three groups reading tests, but was aware of what each student was doing, and assigned additional work as necessary. In another class, a group of students was observed separately with two teachers. The site visitor reported that one would not know the two groups were the same. One teacher monitored the student so that they were much more task persistent and less restless. This teacher was judged to have good responsiveness. The other teacher was rated "so-so."

When a final judgment had been made of each teacher we observed, a numerical rating was assigned to each teacher according to the scheme shown in Table 4-4. We interpret this as a nominal classification scheme.

Table 4-4

CLASSIFICATION SCHEME FOR PROJECT TEACHERS DURING OBSERVATION

Teacher's Implementation of Project	Teacher's Responsiveness		
	Good	So-So	Bad
Well implemented	1	2	3
So-so implemented	4	5	6
Poorly implemented	7	8	9

4.8 Conclusions

Table 4-5 shows the number of teachers given each rating, by project. Despite the lack of detail in the PIP specifications for the instructional program, at least 80% of the ratings were of an acceptable degree of implementation. It would appear that project directors did select the best teachers they could find for the project and that most instructional staff did grasp the PIP philosophy and intent. Only 9 of the 71 responsiveness ratings (13%) were "bad," and only 13 of the 71 ratings (18%) fell into the "poorly implemented" category.

It is not safe to generalize from these ratings because we did not sample teachers. We hoped to get the best and the worst of them. It is clear, however, that the sites and PIPs did differ on their implementation. In Catch-Up we were unable to find any unresponsive teachers. Gloversville Conquest had no well-implemented project teachers. No teacher was found to be implementing R-3 well.

Our interpretation of these variations is dealt with more fully in Volume One. Here, we need only remark that the type of package repre-

Table 4-5

TEACHER RATINGS, BY PROJECT
(In Numbers of Teachers)

Project	Ratings* 1	2	3	4	5	6	7	8	9
Catch-Up									
Bloomington					4				
Brookport				2					
Galax	1	1							
Providence Forge	2								
Wayne City	1				1				
Conquest									
Benton Harbor	1	1		1					
Cleveland	2				1				
Gloversville					1		1	1	2
HIT									
Lexington	1				1				1
Olean	1	1					1	1	
IRIT									
Bloomington	2			1					
Oklahoma City	1				2				
Schenectady		1			1				1
PTR									
Canton				2	2				2
Dallas	2	1			1	2			
R-3									
Charlotte				6	1				
Lake Village							1	2	
Lorain				2	3				
Schenectady					4	1			
Total	14	5	0	14	22	3	3	4	6

*One teacher is rated twice: once for reading and once for mathematics.

sented by all the PIPs, except R-3, could be successful under the conditions of our tryouts. The packages were successful not merely in that teachers mechanically followed directions, but also in that the sense and philosophy of the project was present at the site.

We do not conclude that the package was itself sufficient to introduce superior projects. These sites were fairly intensively monitored. Further, the fact that USOE paid for a conference in Washington doubtless both motivated the participants and gave them information that was not in the PIPs and not in the official modifications, which are attached in Appendix C.*

Insofar as these particular projects are concerned, in this section we have answered two of the issues suggested by the curriculum-referenced analyses. We have defined the PIPs' instructional programs, and we have assessed the degree to which each of our observed teachers was implementing that program.

The next section discusses the actual materials used in the PIP classes and our procedures in ascertaining these materials, and how we assessed the connection between the materials and the MAT.

---

*It should be noted that in addition to the fact that the first year findings served as the basis for revising the PIPs, our second year findings were shared with RMC as the study progressed. These findings have been used as input to the new packages currently being implemented.

5 ANALYSIS OF CURRICULUM AND TEST CONTENT

5.1 Introduction

The first set of field operations--to determine the degree of implementation of the instructional practices--was described in the previous section. In this section, we describe the second set of field operations and associated analyses, which were designed to determine whether the PIP-specified curriculum materials were being used in the field-test projects and whether the tests being used to measure students' performance corresponded with the curriculum to which the students were exposed.

If the skills tested by the MAT were not covered in the PIP-specific curriculum materials used in the projects, we would have no reason to attribute MAT achievement gains to PIP projects. Cursory examination of the MAT in the first year suggested that, in several cases at least, MAT and PIPs were not well matched. Ideally, a detailed investigation of this issue would require:

- Obtaining a record of the lessons that each student covered in each PIP-specified curriculum material.
- Determining the skills required for passing each item on the MAT, for each level administered to students in both pre- and post-tests.
- Determining the skills taught in each lesson in the PIP-specified (and used) curriculum materials.

The test of PIP project effectiveness would then require locating students who failed items in the fall and who subsequently covered PIP-specified curriculum material that was relevant to those items. All other things being equal, well-implemented classes would be expected to show a greater proportion of students answering the failed pretest items correctly on the posttest. If this expectation was not fulfilled, we would have no evidence to claim that PIPs "worked" because a positive association between degree of implementation and outcome would be absent.

However, we were not sure if information could be obtained on individual students. Moreover, we did not know if an analysis of the curriculum materials would be feasible, or how much of a test would be left if we restricted the MAT to only those items that were relevant to the curriculum materials covered. Consequently, we attempted analyses at two levels: an intermediate-level, where we found PIP-specified curriculum materials that were relevant to types of skills included on MAT subtests and that were used with students; a detailed level, where we matched a student with the items he had covered.

5.2 Data Collection Logistics

We attempted first to assess the feasibility of obtaining each student's record of instruction on a daily basis. These data could be aggregated to determine general use of curriculum materials; if the records were detailed enough, they could be used for the individual student-level analysis.

The feasibility of obtaining individual student records was investigated in the fall of the 1975-76 school year. In November the evaluation project analyst visited four tryout projects to determine whether individual records were being maintained on every student in the project (as most of the PIPs themselves required) and, if so, how complete and usable the records were. Most teachers did have available written lesson plans or schedules of instruction (SOIs) with individual assignments for each student.

We already knew that PTR and R-3 projects would not conform to this pattern. PTR projects maintained cumulative records with no association between lessons and dates, and several R-3 projects submitted records without dates. In R-3 we decided to concentrate on the mathematics classes because the curriculum could be more systematically tracked. The R-3 mathematics contracts were approximately the same as assignments, and each student had a record of contracts completed.

During the regular site visits in December, we showed a model SOI to the PIP instructional staff, determined that teachers or tutors were maintaining plans or records containing similar information, and then requested that these schedules be saved for us to examine during our evaluation. During the December visits, each site visitor developed the Dictionary of Core Materials. As described earlier, this dictionary listed materials that the instructional staff claimed to use most frequently and that therefore served as core curriculum. Our guideline for such materials was that they had to be used at least 50% of the time or with at least 50% of the students. Each entry in the dictionary was identified by exact title, publisher's name, date of publication, series, set, level of material used, and other information necessary for us to obtain exact duplicate items or sets of materials.

In January, our site assistants collected a week's accumulation of SOIs from project directors or directly from instructional staff. The site assistant then sent either machine copies or the originals (if teachers did not wish to retain them) directly to SRI. Since abbreviations were used extensively in the notations on the SOIs, each teacher was asked to provide a key to abbreviations used. The key enabled us to match the entries on the schedules with the more complete entries in the dictionaries of core materials.

After examining the SOIs received from all PIP project instructional staff for the week, we saw that the schedules were likely to permit a determination of materials actually covered by the students.* In January we notified teachers not in the observation sample that we would not be collecting any more SOIs, but we requested sample teachers to continue saving their schedules for collection at posttest time. In early March, we asked the sample teachers to give their January through March schedules to our site assistants for transmittal to SRI.

Finally, at the time of spring testing, we asked the sample instructional staff to submit their remaining SOIs for the period through the Friday before test administration. Thus, while we did not collect schedules for the entire period between pre- and post-test, we hoped to have a good idea of the materials used by students from early January to the April test date.

Obtaining SOIs presented some difficulties. About half of the sites had to be telephoned and reminded to turn in some portion of the data. Most were simply slow in sending in all of the schedules for the period from January to posttesting. At the two sites that turned in schedules for only the latter part of the period, site assistants were asked to search teachers' records and submit the missing schedules. By mid-June we had received the SOIs from all sites.

---

*Although SOIs for the lab programs were generally written as plans, not as records of what students actually covered, many teachers noted on these sheets when assignments were not completed. These plans or schedules were proxies, not ideal records of materials covered. They were, however, available and imposed only a slight data collection burden on instructional staff.

5.3 Completeness of the Data

Before using the SOIs in any analyses, we attempted to determine the completeness and quality of the data collected. Table 5-1 shows the number of individual student schedules received for each week. The right-hand column in the table shows the average number of schedules received per week over the number of students with valid spring tests who were in the sampled teachers' classes. Ideally, the numbers should be approximately equal; that is, the number of schedules submitted per week should not vary much from the mean.

Table 5-1 shows that the number of SOIs received for a week was often greater than the number of students in our sample because teachers were asked to send schedules for all students in their groups. Not all students, however, were included in our study. For example, we received schedules for between 45 and 47 students in Providence Forge each week, but there were only 34 students in our evaluation sample. The remaining 13 students were dropped because they exited from the program before posttesting, or had an invalid posttest.

We had expected data for IRIT, for one Catch-Up site, and for PTR to be submitted differently from data sent by the other projects. Because we had included in the tested evaluation group only those IRIT students in the second, or middle, cycle of the program, we were interested in the schedules for only that group. Thus, the period for which we expected schedules in IRIT projects was not January to April, but the approximate ten-week period corresponding to the middle cycle at each site. An additional peculiarity was that each IRIT student had three teachers--one for each of three content areas; thus, we expected three schedules per day per child.

The Catch-Up staff in Galax were making systematic, cumulative records of what each student covered in each curriculum material during the year. We accepted these records rather than collecting, aggregating, and interpreting the daily schedules ourselves.

Finally, as mentioned earlier, in PTR each student's performance was recorded by the tutor on a Record Sheet. Thus, instead of a daily record of instruction, each student had a complete, cumulative record of material covered. Both PTR sites submitted these records for every student in the project, not simply for those in the sampled tutors' group.

Project	Dec 1-5	Dec 8-12	Dec 15-19	Dec 22-26	Dec 29-2	Jan 5-9	Jan 12-16	Jan 19-23	Jan 26-30	Feb 2-6	Feb 9-13	Feb 16-20	Feb 23-27	Mar 1-5	Mar 8-12	Mar 15-19	Mar 22-26	Mar 29-2	Apr 5-9	Apr 12-16	Apr 19-23	Apr 26-30	Mean/N
[illegible]																							
Bloomington						[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	52	[illegible]	52	52	PT			37/57
Brockport						[illegible]	[illegible]	[illegible]	32	[illegible]	[illegible]	[illegible]	31	32	32	31	32	[illegible]	PT				30/27
[illegible]																				PT			29/25
Providence [illegible]						[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	45	45	65	65	65	65	[illegible]	VAC	65	63/46
[illegible] City						[illegible]	28	25	26	[illegible]	25	[illegible]	25	26	26	26	26	26	PT				26/26
Conquest																							
Benton Harbor						74	79	68	68	63	82	82	83	85	57	84	84	55	VAC	84	84	87	76/72
Cleveland						86	86	85	67	95	95	95	88	90	90	90	88	92	PT				87/75
Gloversville																						PT	95/70
[illegible]																							
Lexington						156	156	156	133	128	153	167	150	166	153	25	128	96	69	PT			129/133
Olean						152	164	16[illegible]	166	165	[illegible]	164	166	154	166	164	168	164	PT				165/180
Indi				VACATION	VACATION																		
Bloomington		[illegible]	131			131	125	131	130	131	131	131	[illegible]								PT		113/42
[illegible] City			[illegible]			[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]							PT			37/29
Schenectady							[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]						PT	39/26
[illegible]																							
Canton																				PT			185/81
Dallas																			PT				130/61
[illegible]																							
Charlotte																				PT			287/174
Lake Villa						[illegible]	[illegible]	1	2	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	92	[illegible]	[illegible]	PT		57/126
Lorain																					PT		104/310
Schenectady																						PT	159/118

Key: PT = posttesting, VAC = vacation, D = diagnostic testing; Mean/N = average number of schedules of instruction per week/number of students in sample.

Table 5-1 gives evidence of incomplete data, as discussed below:

- Of the four teachers sampled in Bloomington Catch-Up, one teacher was not asked to keep or retain SOIs for submission to SRI and one turned in records for only three weeks.
- One teacher in Benton harbor Conquest did not begin to submit records until February 2-6. Another did not begin until February 9-13. The third teacher had a few SOIs for January, but did not keep good records until February.
- Two of the teachers in Bloomington IRIT did not submit SOIs for December 8-12 nor for February 23-27.
- The teachers in Schenectady IRIT were not asked to retain SOIs until the sixth week of the second instructional cycle. Only two of these teachers turned in schedules for the remaining four weeks of the cycle.
- Charlotte and Schenectady R-3 project teachers did not include dates on their SOIs.
- The only R-3 mathematics teacher in Lake Village did not turn in many schedules prior to the second week of February.
- One teacher in Lorain R-3 did not submit any schedules, and the other teacher did not include dates on the schedules.

The Gloversville and Olean SOIs allowed some major findings about the implementation of the respective Conquest and HIT projects in those sites. The schedules received from Gloversville contained no weekly information because they were not kept on a daily or weekly basis. This was our first indication that students there did not attend the lab on a regular, five-days-per-week basis, but rather on an irregular basis as assigned by their classroom teachers. Entries on the schedules were spotty or incomplete; presumably the regular classroom teacher rather than the PIP clinician maintained records of student progress. In any case, the number of schedules received was larger than the number of students in our sample; the reason is that the students in two of the teachers' classes were excluded from our evaluation sample because they had received so little of the Conquest treatment that it was considered unfair to evaluate the PIP on the basis of their performance.

In Olean, we received records for 233 students thought to be in the project on the specified four-days per week. However, 130 of these students were scheduled in the HIT classes only every other week. That is, one or two teachers alternated student groups each week. We had not detected this scheduling pattern in our earlier interviews and observations. For the 185 students with valid test data, an average of 165 records were received.

Although Table 5-i reveals a number of irregularities, we received from most teachers in most sites a fairly complete and reliable record of the materials covered. Only one entry gave us reason to question the validity of the data: Although March 15-19 was Easter vacation for Lexington, one teacher there submitted schedules for 25 students. Perhaps she had made plans for students for that week before realizing that it was a vacation period, but the quality of the entries on these schedules and our knowledge of the teacher's organizational style led us to believe otherwise. Exhibit 5-1 shows four schedules of instruction received from the field. They vary in completeness and format, but three of them (with their keys) contain usable information. Schedule A (on Exhibit 5-1) does not contain adequate information about curriculum materials. The teacher who prepared this schedule also received low ratings on implementation and responsiveness during our interview and observation visits. This example reflects a general observation we made about the data. SOIs that were incomplete or of poor quality were submitted by the same instructional staff in our sample who also received the lowest ratings on implementation of the PIP instructional program. Thus, quality of curriculum information on the students is confounded with degree of PIP implementation. Any possible bias in the remainder of the analyses would lie with the well-implemented instructional groups or the "good" teachers because more of their data were usable in our analyses.

In general, we believe that enough SOIs were received for the entire sample group for the entire period and that they provide fairly reliable information about what was being covered in the PIP projects. Although we received information for about 1700 students, clearly we did not have a random sample of student lesson assignments. Nevertheless, SOIs were a fairly direct measure of what curriculum was used in PIP classes and were clearly superior to interviews and observations alone.

5.4 Congruence Between PIP-Specified Curriculum Materials and Materials Used in the Field-Test Projects

For each project, we compiled a list of all equipment and curriculum materials that appeared anywhere on the SOIs. The number of titles

EXHIBIT 5-1 EXAMPLES OF SCHEDULES OF INSTRUCTION RECEIVED FROM FIELD-TEST SITES

A.

Mrs. [illegible] - - Lexington [illegible]

SCHEDULE OF INSTRUCTION

Week of 3 (Mo.) 22 (day) through 26 (day)

	MONDAY	TUESDAY	WEDNESDAY	THURSDAY	FRIDAY
Olive, Fred	Flashcards Division Worksheets	Flashcards Division Worksheets	Flashcards Division Worksheets	Flashcards Division Worksheets	
Beat, Dee Dee	Flashcards Division Worksheets	Flashcards Division Worksheets	Flashcards Division Worksheets	Flashcards Division Worksheets	No Lesson Information
Michaels, William	Flashcards Division Worksheets	Flashcards Division Worksheets	Flashcards Division Worksheets	Flashcards Division Worksheets	
Rice, Harry	Flashcards Division Worksheets	Flashcards Division Worksheets	Flashcards Division Worksheets	Flashcards Division Worksheets	

B.

Clinician- Mrs. Pam Smith Grade 2

Name Mary White City - Benton Harbor

School- Fair Plain Northeast Conquest

1976 Daily Record Sheet

Date	Book I Sullivan Reading	Consonant Sounds Phonovisual	Comprehension	Vocabulary	Word Atta
Feb 2	p. 17-18	Phonov. "j" Sound-Study		Dolch Words Lang. Master	
Feb 4	p. 19, 20	Phonov. Worksht. p. 17			
Feb 9	p. 21	Phonov. "g" Sound-Study	SRA-1A Listen Story	Dolch "A" Words	
Feb 10	p. 22, 23	Phonov. Worksht. p. 18	Skilltext BLA p. 46		
Feb 11	p. 24	Clues p. 24	Martin L. King Black History Wk. Study		
Feb 13	Game Day				
Feb 16	p. 25, 26	Clues p. 23	SRA-1A Listen. Story		

Good Everything we Need

C.

Teacher - John Doe
Bloomington, I.R.I.T.

Name Ann Read Dec. 15

Mon.

Some lessons OK
Others are incomplete

Absent

Tue.

Absent

Wed.

1 SRA [illegible]

2 Getting the Main Idea A 7,8,9 [illegible]

Thur.

1 You did the wrong SRA! See me

2 Read DTC - Chap 1+2

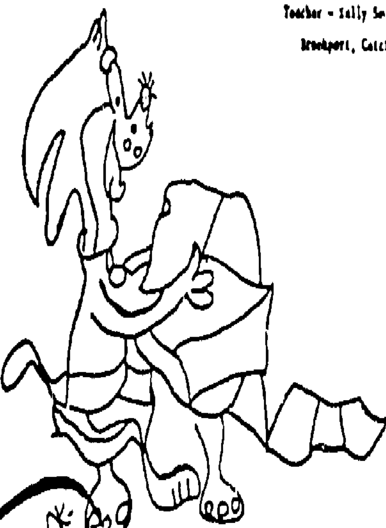
3 Written exercises for Chap 2 (See me)

Fri.

Class lesson on main idea

D.

Teacher - Sally Smith
Brockport, Catch-up



Some lessons OK
others are incomplete
need page numbers

A record for Susie Student
Week of Jan 19-23

	Reading	Math
Monday	34.4 (7,9) [illegible]	Mult Records 6,7's
Tuesday	34-15 wkshts 40N [illegible] p112	Tutor 6's, 7's
Wednesday	MAC Curris phonics wkbk [illegible] p182	Sullivan wkbk 26
Thursday	38-3 [illegible] The Main Idea	Sullivan wkbk 26
Friday	38-3 Main Idea	Mult Records 8's, 9's

175

ranged from 2 for Dallas PTR to 65 for Schenectady IRIT. With each list we addressed the question: Did the project staff use the PIP-specified materials? The problem (as reported in Section 4) was that the materials specifications in the PIPs were often ambiguous. We decided to list all the materials mentioned in each PIP and designate them as "specified" materials, with a distinction made between "core" materials and "supplementary" materials.* These were the "PIP-specified materials" against which we compared the list of materials generated by aggregating across all SOIs collected from the sampled teachers.

Our findings from this comparison were that:

- Some overlap existed between what was PIP-specified and what was used; that is, some of the PIP-specified materials were used.
- Many of the materials specified in the PIP were not used.
- Teachers used a surprisingly large number of materials that had not been listed in the PIPs.

Tables 5-2 through 5-6 summarize our findings about the instructional materials used at each project site. In each table, the left-hand column indicates the number of PIP-specified materials used and the percentage of the specified instructional curriculum this represented for each project. The right-hand column indicates the number of non-PIP-specified materials used and the percentage of the total instructional materials this represented for each project. In addition, a separate entry shows the number of materials, either specified or nonspecified, used by all the sites with the same PIP program. This number is also included in each project's figures. These data show that not every site using the same PIP used the same combination of specified materials, nor did sites reject the same combination of specified

---

*We could not always determine which materials were "specified." When two materials were called by the same name, we counted both as specified materials. For example, McGraw-Hill Programmed Reading and Sullivan Reading Program were both referred to as "Sullivan" by teachers. (McGraw-Hill Publishing Company had bought out BRL, publishers of the Sullivan Reading Program, and had issued a completely revised series.) Another difficulty was with the Catch-Up PIP, which referred to the SRA Math Kit. There was nothing by that title, although there was an SRA Math Learning System and an SRA Diagnosis: An Instructional Aid-Mathematics which could have fit the material described, and therefore we called both of them specified materials.

Table 5-2

SPECIFIED AND UNSPECIFIED MATERIALS
USED AT CATCH-UP SITES

a. Reading

Reading Materials, by Project	Specified: Used N	Specified: Used % of Total Specified	Specified: Not Used N	Specified: Not Used % of Total Specified	Not Specified and Used N	Not Specified and Used % of All Materials Used at Each Site*
Core materials						
Bloomington	6	100.0%	0	0 %	33	75.0%
Brookport	2	33.3	4	66.7	23	79.3
Galax	3	50.0	3	50.0	17	77.3
Providence Forge	3	50.0	3	50.0	12	63.2
Wayne City	4	66.7	2	32.3	16	72.7
Used by all sites	1	16.7	--	--	2†	2.0
Total no. core materials specified = 6‡	Total no. core used = 6§				Total no. unspecified materials = 84	
Supplementary materials						
Bloomington	5	35.7%	9	64.3%	--	--
Brookport	4	28.6	10	71.4	--	--
Galax	2	14.3	12	85.7	--	--
Providence Forge	4	28.6	10	71.4	--	--
Wayne City	2	14.3	12	85.7	--	--
Used by all sites	--		--	--	--	--
Total no. supplementary materials specified = 14†	Total no. supp. used = 11‡					

*The ratio between the number of "not specified and used" materials and the total number of materials used at a site. (The total includes all materials used at a site, both specified--core and supplementary--and unspecified.)

†Credit for one material each is given to teacher-made materials, games, and free reading even though many unidentified materials may have been used within each category.

‡Includes one type of equipment.

§Equipment use was not always specified in conjunction with use of materials mentioned in the schedules of instruction.

Table 5-2 (Concluded)

b. Math

Math Materials, by Project	Specified				Not Specified and Used	
	Used		Not Used			
	N	% of Total Specified	N	% of Total Specified	N	% of All Materials Used at Each Site
Core materials						
Bloomington	3	100.0%	0	0 %	6	54.5%
Brookport	3	100.0	0	0	14	77.8
Galax	1	33.3	2	66.7	10	83.3
Providence Forge	2	66.7	1	33.3	7	70.0
Wayne City	3	100.0	0	0	15	78.9
Used by all sites	1	33.3	--	--	1*	2.1
Total no. core materials specified = 3	Total no. core used = 3				Total no. unspecified materials = 43	
Supplementary materials						
Bloomington	2	40.0%	3	60.0%	--	--
Brookport	1	20.0	4	80.0	--	--
Galax	1	20.0	4	80.0	--	--
Providence Forge	1	20.0	4	80.0	--	--
Wayne City	1	20.0	4	80.0	--	--
Used by all sites	1	20.0	--	--	--	--
Total no. supplementary materials specified = 5+	Total no. supp. used = 2					

*Credit for one material each is given to teacher-made materials, games, and free reading even though many unidentified materials may have been used within each category.

+The hardware/software package did not clearly indicate whether one of the materials referred to was the SRA Math Learning System or the SRA Mathematics Diagnosis; therefore, each was considered a specified material.

Table 5-3

SPECIFIED AND UNSPECIFIED MATERIALS
USED AT CONQUEST SITES

Reading Materials by Project	Specified: Used N	Specified: Used % of Total Specified	Specified: Not Used N	Specified: Not Used % of Total Specified	Not Specified and Used N	Not Specified and Used % of All Materials Used at Each Site
Core materials						
Benton Harbor	16	61.5%	10	38.5%	36	66.7%
Cleveland	16	61.5	10	38.5	40	69.0
Gloversville	8	30.8	18	69.2	55	84.6
Used by all sites	5	19.2	--	--	3*	2.1
Total no. core materials specified = 26†	Total no. core used = 20				Total no. unspecified materials = 119*	
Supplementary materials						
Benton Harbor	2	33.3%	4	66.7%	--	--
Cleveland	2	33.3	4	66.7	--	--
Gloversville	2	33.3	4	66.7	--	--
Used by all sites	1	16.7	--		--	--
Total no. supplementary materials specified = 6	Total no. supp. used = 3					

*Credit for one material each is given to teacher-made materials, games, and free reading even though many unidentified materials may have been used within each category.

†Includes six types of equipment. Three materials mentioned in the hardware/software package are contained in the Webster Classroom Reading Clinic but are counted as separate materials.

Table 5-4

SPECIFIED AND UNSPECIFIED MATERIALS
USED AT HIT SITES

Core Materials by Project	Specified: Used N	Specified: Used % of Total Specified	Specified: Not Used N	Specified: Not Used % of Total Specified	Not Specified and Used N	Not Specified and Used % of All Materials Used at Each Site
Core materials--reading						
Lexington	3	60.0%	2	40.0%	6*	66.6%
Olean	3	60.0	2	40.0	3	50.0
Used by all sites	1	20.0	--	--	3	7.7
Total no. core materials specified = 5†	Total no. core used = 5				Total no. unspecified materials = 8*	
Core materials--math						
Lexington	3	25.0%	9	75.0%	3	50.0%
Olean	3	25.0	9	75.0	3	50.0
Used by all sites	2	0	--	--	--	--
Total no. core materials specified = 11	Total no. core used = 4				Total no. unspecified materials = 6	

*Credit for one material each is given to teacher-made materials, games, and free reading even though many unidentified materials may have been used within each category.

†One material (Conquests in Reading) was included in the hardware/software package by mistake but was used in Lexington because they had problems obtaining the Sullivan Programmed Reading material.

Table 5-5

SPECIFIED AND UNSPECIFIED MATERIALS
USED AT IRIT SITES

Reading Materials, by Project	Specified: Used N	Specified: Used % of Total Specified	Specified: Not Used N	Specified: Not Used % of Total Specified	Not Specified and Used N	Not Specified and Used % of All Materials Used at Each Site
Core materials						
Bloomington	10	18.2%	45	81.8%	34	63.0%
Oklahoma City	7	12.7	48	87.3	24	63.2
Schenectady	5	9.1	50	90.9	50	76.9
Used by all sites	2	3.6	--	--	3*	2.5
Total no. core materials specified = 55†	Total no. core used = 12‡				Total no. unspecified materials = 92*	
Supplementary materials						
Bloomington	10	12.2%	72	87.8%	--	--
Oklahoma City	7	8.5	75	91.5	--	--
Schenectady	10	12.2	72	87.8	--	--
Used by all sites	5§	6.1	--	--	--	--
Total no. supplementary materials specified = 82**	Total no. supp. used = 17‡					

*Credit for one material each is given to teacher-made materials, games, and free reading even though many unidentified materials may have been used within each category.

†Includes one type of equipment.

‡Equipment use was not always specified in conjunction with use of materials mentioned in the schedules of instruction.

§All are equipment.

**Includes 13 types of equipment.

Table 5-6

SPECIFIED AND UNSPECIFIED MATERIALS
USED AT R-3 SITES

Math Materials,* by Project	Specified				Not Specified and Used	
	Used		Not Used			
	N	% of Total Specified	N	% of Total Specified	N	% of All Materials Used at Each Site
Core materials						
Charlotte	2	25.0%	6	25.0%	19	82.6%
Lake Village	2	25.0	6	75.0	3	42.9
Lorain	3	37.5	5	62.5	6	60.0
Schenectady	2	25.0	6	75.0	7	77.8
Used by all sites	1	12.5	--	--	--	--
Total no. core materials specified = 8	Total no. core used = 4				Total no. unspecified materials = 30	
Supplementary materials						
Charlotte	2	12.5%	14	87.5%	--	--
Lake Village	2	12.5	14	87.5	--	--
Lorain	1	8.3	15	91.7	--	--
Schenectady	0	0	16	100.0	--	--
Used by all sites	0	0	--		--	--
Total no. supplementary materials specified = 24	Total no. supp. used = 3					

*The R-3 program also includes reading and social studies instruction, but we were able to adequately assess only the materials used in the math classes.

materials. Some projects used more unspecified materials than other sites (varying from about 43% to 85% of their materials), and some PIPs were associated with more use of nonspecified materials than were other PIPs.

Data for PTR projects are not tabled, because they are so easy to describe. The PIP required only that PTR projects employ a tutoring manual that complemented the basal reading text used in the regular classroom. A supplementary material, the Alphabet Skills Book, was also mentioned in the PIP. Both materials were used in Canton and Dallas. (In Canton, however, where students enter at the first grade level without kindergarten, no basal readers had been used before the PIP arrived. These texts had to be obtained before the PIP-specified materials could be used as intended.)

Among other PIPs, Catch-Up projects appeared to use the greatest percent of specified materials for reading instruction, with Bloomington including 100% of such materials in their curriculum. Even with the overlap of materials among Catch-Up projects, only one specified reading material was used at all Catch-Up sites. All Catch-Up projects used many unspecified materials for both reading and math, with each project using a different set of such materials.

With the exception of Gloversville, Conquest sites used about 60% of the core materials specified in the PIPs. A large number of unspecified materials were added at every site.

The 121 reading materials used in IRIT corresponded even less closely to the specified lists, and only two of 55 specified core materials were used in common across the sites. More unspecified materials were used in IRIT than in either of the other two lab programs.

The Catch-Up, Conquest, and IRIT PIPs all had long lists of materials. Except perhaps in the Catch-Up PIP, the original Hardware/Software Packet did not clearly state which materials were core. In all three PIPs, various materials were suggested and instructional staff were encouraged to assemble a variety of materials and to individualize instruction by providing students with materials suited to their needs. The Catch-Up PIP stated that each teacher should have his own funds to purchase the materials he liked and should use materials in any way he deemed practical.

Like the lab programs, the R-3 PIP encouraged use of a variety of materials. Even more than the others, it encouraged instructors to seek out materials that were likely to motivate individual students. As expected, the R-3 projects did use unspecified materials and did not use a large percent of the recommended materials.

Except for PTR and to some extent HIT projects, little congruence was found between the specified (core and supplementary) materials and the materials actually used in the project. This finding is surprising for those who expected the instructional programs in sites with the same PIP to be the same in terms of materials and equipment used. Insofar as comparison of titles permitted us to tell, packages written in the manner of five of the original six PIPs would not promote the use of a common set of curriculum materials in new sites.

We cannot say, however, that projects deviated from PIP instructions. Although our comparison of titles shows that PIPs (except PTR) failed to promote the use of exactly those materials and only those materials recommended, this was probably not a violation of RMC's intent in developing the PIPs because Catch-Up, IRIT, R-3, and Conquest PIPs encouraged teacher discretion in choosing materials.

We are not naive enough to assume that teachers had to use exactly the same materials in order to implement effective curriculum.* We know that, especially when individualization is required, different combinations of texts, equipment, and other teaching materials were necessary to carry out the intent of the curriculum. Nevertheless, PIPs were presumed to present enough information about pedagogical philosophy and skill emphases to (1) permit teachers to ascertain the essence of the effective curriculum of the original project and (2) promote the use of materials incorporating the intended skills lessons in the intended manner. With the help of reading curriculum specialists, our next analysis enabled us to assess whether, in spite of the variety of materials used, the PIP projects nevertheless adhered to the curriculum intent of the PIPs and covered the same skills in a similar manner.

*Although in this report we often use "curriculum" to mean the materials and equipment to which students are exposed, we use the term here in its general meaning as "planned learning experiences encountered by students," a definition that includes the instructional philosophy as well as the knowledge and skills covered in the materials.

5.5 Analysis of the Core Curriculum at Each Project

Before we could match project curricula with PIP-specified curricula, we had to determine what materials were being used as the core of the instructional program. We examined the SOIs again for the materials used most often. (The list of all materials used by each project was too gross for this analysis because once-used titles also appeared on this list.)

After determining which materials were used most frequently by project students, we called on reading curriculum specialists to help us describe the skills emphasized in each.* This gross analysis was conducted by means of a skills checklist. As stated earlier, our general intent was to determine the relevance of the MAT to the PIP curriculum. We also intended to use the skills analysis to determine whether, even when different sets of materials were used, the project staffs understood and implemented the PIP-intended curriculum.

5.5.1 Procedures

Because we received SOIs for approximately 1700 students, time and effort dictated that information on the use of materials not be tabulated from every schedule. Instead, we sampled schedules by PIP, project, grade, and teacher/tutor. A sample of five students was picked from each grade at each site.

The grades were grouped roughly according to the MAT battery that students received in the spring because we expected to compare the general skills known to be covered in the materials with the general skills tested on the MAT. The only exception was that we grouped first and second grades, even though students in these grades took different tests. Groupings were as follows:

- Grades 1 and 2
- Grades 3 and 4

---

*Consultants were: Ms. Patricia Bixler, former reading curriculum coordinator, San Mateo County Schools (California), currently principal of Knolls School in San Mateo; and Dr. Arlene Bonnie Tenenbaum, former SRI consultant, currently evaluation specialist for Cupertino School District (California).

- Grades 5 and 6
- Grades 7, 8, and 9.

Besides grade level, individual teachers were likely to affect which materials were designated as core. At each site, the five students sampled were drawn so as to be distributed across the teachers in that site and across the grade levels for which each teacher had responsibility. For example, in the Bloomington Catch-Up project, our observation sample had three teachers. Teacher A had only first graders; teachers B and C had only second graders. The sample included the following students:

- Two first grade students from teacher A
- Two second grade students from teacher B
- One second grade student from teacher C.

To tabulate the list of core materials for each project, we used the following procedures: One entry was recorded for each material used by a student each day. If the student received two lessons in a material on a single day, only one entry was recorded for that material.

The frequency data is shown in Tables E-1 through E-4 in Appendix E. The data reflect the full period for which schedules were received, that is, from January to posttesting, for all PIPs except IRIT. IRIT's SOIs cover the students enrolled in the second cycle only. Because neither PTR nor R-3 sites submitted SOIs on a daily or weekly basis, the procedures could not be carried out for these two programs.

The ten materials most frequently used by each project during this period before posttesting were tabulated. For each PIP, Tables E-1 through E-4 show the ten materials used as the core of the curriculum by each project. The second column in each table indicates if the material was specified as core or supplementary material in the PIP.

The individual skills into which the materials could be most comfortably categorized were as follows:

- Recognition of Sounds and Letter Recognition--Designed to teach letter-sound correspondence and visual discrimination of letters.
- Decoding--Sometimes used synonymously with phonics (associating a letter or combination of letters with a sound and applying such knowledge in identifying words). Lessons

are designed to teach word identification and converting print into speech. Included are pronunciation and associating a group of letters making up a word with the sounds in its spoken counterpart.

- Structural Analysis--Entails looking at words to locate parts of them (e.g., syllables, prefixes, suffixes, special endings, root words). Structural analysis may be used in conjunction with phonics (phonetic analysis) and context clues to identify a word.

- Vocabulary--Entails gaining knowledge of the meaning of a word and learning to recognize it in print. (Words in the curriculum materials were assumed to increase in complexity and to decrease in frequency of exposure as grade level increased.)

- Antonyms and Synonyms--Used to increase vocabulary.

- Comprehension--Entails understanding the meaning of a written word, a written sentence, or a written passage of one paragraph or more. Responding to questions and acting on the information read are included.

Recognition of sounds and letter recognition are beginning reading skills and are usually covered in the first grade. These skills are included only on the MAT Primer, which is the battery for entering first graders. The next two skills, decoding and structural analysis, are also beginning reading skills and are usually not included on norm-referenced tests for students above the first grade. Vocabulary can cover a huge range, depending on the complexity of the words and the frequency of exposure in appropriate contexts. Vocabulary items are included on every level of the MAT from Primer to Advanced. Reading comprehension also covers a range of skills and interactions of skills that have never been satisfactorily understood. Generally tests of reading comprehension include passages of increasing length and have increasingly complex vocabulary and syntax for students in first grade and above. Reading subtests on the Primary I, Primary II, Elementary, Intermediate, and Advanced MAT all test reading comprehension.

Because we did not tabulate information from the SOIs about the levels in the curriculum series at which the students were performing, we asked the reading curriculum specialists for each grade group to check only the skills appropriate to the given grade level or below. That is, we asked the specialists to assume--especially when analyzing curriculum series intended for kindergarten through sixth grade levels--that none of the students would be covering materials above the average level of difficulty for their grade. Several other assumptions were also necessary:

- That materials were used as intended by publishers or manufacturers for full effectiveness (e.g., that both the auditory and visual components of the Auto-Vance machine were used as designed).
- That teachers provided the necessary instruction for each student as outlined in published manuals and as suggested in the PIPs.
- That exposure really meant covering the material adequately enough to learn the skills.

5.5.2 Analysis

Tables E-1 through E-4 in Appendix E show the most frequently used materials and a checklist of the skills covered in those materials. A pattern is revealed in the Frequency of Use columns at grade levels where projects with the same PIP can be compared: Each site shows a different set of most frequently used materials. In the most extreme case, seen in Table E-4b (individualized reading instruction) for the IRIT projects, the frequently used materials form virtually nonoverlapping sets.

The tables show that when project lists overlap, they usually overlap on those materials specified in the PIP as core materials. Consequently, although the core sets of materials differed among sites, we have a slight indication that the instructional staff understood the core of the curriculum intended by the PIPs. Since the reading materials used in every lab project did cover the entire range of skills, the skills checklist is disappointing as an indicator of the degree to which projects implemented the curriculum. However, we know from interviews and informal observations during site visits that project staff generally understood the skill emphases intended in the PIP, even when they did not implement those emphases.

The reading skills checklists do not reveal any differential relevance of the MAT battery to the curricula, except perhaps for HIT; in HIT Lexington the more advanced levels of the MAT, requiring reading comprehension, would be fairly irrelevant to the curriculum. A skills analysis performed on the curriculum materials in the absence of information about levels, and about lessons at which students were actually placed within them, was too weak to reveal relevance to emphases in the MAT. The skills checklist proved disappointing because from our interviews and class observations we had the strong impression (1) that although the materials being used contained lessons on reading comprehension skills, such as those emphasized on MAT Primary II through Advanced tests, teachers placed little emphasis on these higher-level reading skills and (2) that few students were studying anything but remedial phonics, decoding, structural analysis, and vocabulary.

Table E-1 shows considerable variation among most-used reading materials, but shows that all Catch-Up sites used the Random House Criterion Reading kit as one of their core materials. This diagnostic, skills-testing kit was keyed to lessons in only a few of the materials specified in the Hardware/Software Packet, and the Project Director's Manual did not clearly describe how teachers should index their teaching materials to the skill areas in the Random House kit. Teachers in every site except Galax attempted to choose materials that covered the skill areas in the Random House series and attempted to index the lessons, games, books, or worksheets accordingly.

The Catch-Up PIP described an eclectic approach to reading instruction and provided little guidance on what to teach or on how to teach particular skills. Rather, the PIP encouraged teachers to exercise judgment in choice of materials and suggested methods for providing frequent success experiences and praise for the students. This pedagogical philosophy was understood by project staffs. For teaching reading and remediating reading difficulties, the PIP recommended selecting what appears to be best for each child from among a variety of materials and equipment; emphasis on phonics was inferred from the specification of the Random House Criterion Reading Kit.

In Wayne City, the Random House series was taken so seriously for determining skill coverage essential to Catch-Up that staff created a "Core File" (shown as a most frequently used material in Table E-1. This file contained individual worksheets and pages removed from a variety of published series and kits and filed according to the skills index specified by Random House. The project director at Wayne City purchased the Fountain Valley diagnostic/prescriptive kit to aid teachers in identifying materials that presented lessons keyed to the skill deficiency areas. As shown in Table E-1, Catch-Up sites also had a major portion of their most frequently used mathematics materials in common with the PIP-recommended core.

Except for Gloversville, where staff were implementing the Wisconsin IGE program rather than the Conquest program, the Conquest projects shown in Table E-2 used the PIP-recommended core materials more frequently than did other PIP projects. Each used some materials focusing on reading comprehension skills, but the primary emphasis was on preliminary reading skills and vocabulary development. Instructional staff in Conquest projects understood and implemented the curriculum intended in the PIP.

In HIT (Table E-3), Remedial Reading Drills was used quite frequently in both Lexington and Olean reading centers. Lexington also employed the PIP-specified Stories of the Inner City, but Olean did not. During the first year RMC reported that Stories of the Inner City was mistakenly included in the HIT PIP. (It had been transposed from the Conquest PIP.) For the second school year, the Lexington project staff made several attempts to get the Sullivan company representative to deliver the materials used in the original HIT site, but were unsuccessful. They continued to use Stories of the Inner City to supplement the heavy core curriculum emphasis on phonics drill work. Olean had used the recommended Sullivan materials since the beginning of the field test, but staff felt Sullivan was too limited in its focus on phonics and pronunciation. In this site, staff added more reading comprehension materials to their curriculum than the PIP intended.

Of all the PIPs, IRIT (Table E-4) had what one reading specialist called "the most well-balanced reading curriculum." The original PIP explicitly covered decoding, vocabulary and comprehension, and individualized reading for comprehension and enjoyment. More than the other two lab programs (especially more than Conquest), IRIT relied on the teachers to select the materials they thought would best teach the skills to their students.

The IRIT PIP designated 51 core materials and 71 supplementary materials. The sheer quantity of core materials indicated that the intent was to specify a very loose base from which teachers were free to vary. Materials listed covered a variety of vocabulary and phonics materials, reading comprehension workbooks, storybooks for fun reading, and audiovisual materials for motivation and enjoyment. In general, descriptive information of recommended materials and their use was not provided by the PIP. In addition, many of the recommended materials covered the same area of instruction (e.g., several readers were recommended); that teachers were to choose from among the materials was only implied. This confusion has been eliminated by the revised IRIT PIP, which clearly specifies that teachers choose one or two materials from each category of recommended materials. The revised IRIT also includes more detailed explanations of skill coverage in each category. The recommendation in the Hardware/Software Packet (now called Materials/Equipment Catalogue) that teachers choose materials different from those used in the regular classroom remains in the revised PIP.

Although the IRIT field-test projects did not use many of the specified materials, the spirit of the program seems to have been carried out. All three sites adequately covered the areas of phonics, vocabulary and comprehension, and individualized reading. The projects placed slightly more emphasis on spelling, comprehension, vocabulary development, and phonics than did the PIP. Bloomington and Schenectady used more criterion-referenced skills testing, and Oklahoma City added typing as an application of language arts. In the main, however, the same general skill areas were emphasized.

In summary, several points can be made about the congruence of the core curricula in the projects and in the PIPs and the relevance of those core curricula to the MAT:

- The skills checklists suggested that the MAT was perfectly appropriate for measuring the specified-and-used curriculum. We know from other observations, however, that the curriculum materials contained a whole range of skills (e.g., reading comprehension skills) that few students covered. Moreover, unless one knows which lessons are being studied within the curriculum materials, one does not have much information about skill emphases. Knowledge of the specific skills that individual students have been studying is necessary for determining whether the MAT is a valid measure of project curriculum.
- None of the PIPs--except PTR and, to a large extent, HIT--contained information about exactly which materials were responsible (along with effective teaching) for the effectiveness of the exemplary program in its original site. This is not a criticism of the PIPs as a communication device. PIPs carried the message (from the original project staff and RMC analysts) that the use of exactly the same materials was not necessary. Such a message, however, was devastating to the PIPs as a replication device that promised to prescribe the conditions required for achievement gains.
- Under the Title III grants, which required that adherence to the PIPs be monitored, project staffs sought to understand the skill emphases and other aspects of curricular philosophy communicated in the PIPs. They generally understood the curricular philosophy but could infer skill emphases more easily when materials specifications were clear, core and supplementary materials were distinguished, and the recommended list was limited to a few materials.

> Most project staffs decided to endorse and implement the skill emphases they inferred. Some (e.g., Gloversville Conquest and Olean HIT) decided to reject or modify them. When open-ended recommendations were made in the PIPs (e.g., to acquire materials that would enable them to teach individual students better), staff often searched for more guidance and structure than the original PIPs provided.

Our observations and interviews with teachers indicated that they were likely to use already-familiar materials if given a choice between those and others that were designed to accomplish approximately the same objectives. In addition, when teachers were unfamiliar with curriculum materials, they were more inclined to use them if they could be given convincing reasons for doing so, were given information on how to use them, or, best of all, had time to familiarize themselves with the materials during pre-program, in-service training. These observations resulted in revisions in the Materials/Equipment Catalogue and some additions in a new Training Manual. The revised PIPs for the three lab programs include more information on the purposes, advantages, and disadvantages of each material. The R-3 and lab PIPs, however, still allow teachers the freedom to choose among the many materials that were present at the original project site.

Our observations that teachers used already-familiar materials made us question whether the PIP projects were innovations at each site, whether the PIPs had influenced teachers to adopt new materials suggested in the packages, and whether the PIP project curricula at each site were different from the regular school curricula.

5.6 The Regular Classroom Curriculum

Although the evaluation was designed to enable us to attribute effects to the PIP projects, achievement gains clearly cannot be attributed solely to the projects. For all projects except R-3, students spent most of their instructional time during the year in their regular classrooms; thus, posttest scores at the end of the school year were affected by both the special project and the regular instructional program. An important consideration for our interpretation of any achievement gains was the nature of the alternative curriculum--that is, the curriculum in the non-PIP classrooms from which PIP students were sent.

Although we wanted to know about the reading and math skills the students were learning in their classrooms, we did not have the massive resources for a thorough study of the regular curriculum. However, to achieve a rough idea of the alternative explanation for changes in achievement, regular classroom teachers or school principals were asked to list their core reading and math materials. We thus achieved a general idea of the skills emphasized in regular classrooms and attempted to examine the relationship of these skills to the PIP curricula.

As discussed, the PIP programs, except for R-3, were supplementary programs. Catch-Up, Conquest, and PTR were daily "pull-out" programs that were not supposed to replace participation in regular classroom reading instruction. IRIT students spent approximately three hours in the IRIT classrooms, during which they could have missed their regular reading instruction, but they participated in the IRIT program for only one cycle (i.e., a 10- or 11-week period).

A review of non-PIP reading and math curricula revealed that, although the materials varied, the skills included in the PIP curriculum at each site were also covered in the regular classroom. A single basal reader was central to almost every elementary level, non-PIP reading curriculum; these basal readers were sometimes also used in the PIP classes to ensure that the skills learned in the projects were appropriate to the performance required in the regular classroom. Gloversville Conquest was an exception. There the curricula in both the regular classroom and the lab were the same; they were based on the Individually Guided Education program developed at the Wisconsin R&D center.

The pedagogical philosophy in reading was the same from classroom to lab in most sites. Both the PIP projects and the regular classroom emphasized phonics. Variation between project and regular classes may have occurred in the sequence and manner in which new skills were introduced, with somewhat greater emphasis on reading comprehension in the basal readers and an increased degree of attention to individual needs given in the projects. It seemed likely that some of the unspecified materials being used by project teachers were those materials that they had used previously in the regular classroom.

The curricula of the HIT and R-3 programs were different from the other four PIP projects and require additional comment. The HIT curriculum, as outlined earlier, centered upon remedial phonics. HIT students did little work on the reading comprehension skills being practiced by (though not specifically taught to) other students at their grade level. Reading achievement gains could probably be attributed to the HIT curriculum except that phonics skills were not tested on the test batteries for students in grades 7-9. It is more likely that practice in reading comprehension in their regular classrooms would have helped PIP participants most on the MAT. The R-3 program integrated the reading, math, and social studies skills of an entire grade level at a participating school. Teachers were expected to intersperse the recommended R-3 instructional techniques among more familiar methods, thus encouraging students to respond more enthusiastically to the curriculum and to achieve greater academic gains. Technically, therefore, any achievement gains could be attributed to the curriculum of the R-3 projects because the entire school year experience of the students was their R-3 program. On the other hand, we do not know whether students would have performed differently in the absence of the R-3 program.

Clearly then, except for HIT, projects did not have a curriculum that was significantly different from the regular classroom curriculum. Although the IRIT and Catch-Up PIPs recommended using materials and equipment that were different from those used in the classroom, they did not intend differences in skill coverage; they meant only to provide lesson variety. Thus, for all but HIT, it would be difficult to separate the effectiveness of the PIP from that of the regular curriculum.

Failure of PIP students to make gains when their project teachers used the recommended materials and followed the PIP instructional style (assuming the test is appropriate to the skills covered) would indicate that the PIP had failed. Success of PIP students in achieving gains, on the other hand, would have to be attributed to both the PIP and the regular curriculum because we cannot separate PIP effects from the effects of the regular classes which PIPs are designed to supplement.

5.7 Detailed Correspondence Between the MAT and Fourth and Eighth Grade Curriculum

The analysis reported in the preceding sections are at a fairly gross level of detail. In our view a much better analysis would be to match our data on lesson plans with the MAT item scores. Surely a convincing analysis of PIP project failure would be that items failed in the fall and known to be covered during instruction, were not passed in the spring. As already noted, one could not show success if the items were passed unless a way was found to argue that the regular school curriculum did not cover the items.

The outcome of an argument based on known-to-be covered items seem to us to be so compelling that we attempted an analysis which matched curriculum information and item scores on our fourth and eighth grade children, where we gave the same test fall and spring.

The results were rewarding in that we found so few items and children for analysis, we felt confident that we had partially explained why the original project results were not replicated: the MAT did not test what these compensatory reading teachers were doing.

The results were disappointing in that we did not feel that we could report any formal analyses on such a thin data base. However, we report the steps of our procedure in Appendix G, for those who may wish to try an analysis at this level of detail.

Our curriculum analyses have shown that the MAT was not especially relevant to the PIP project materials. The question naturally arises, were the original validating tests any better? The next section addresses this issue.

5.8 Tests Used to Validate Original Programs Compared with the PIP Curriculum and the MAT

Ostensibly, we set out to evaluate the effectiveness of the PIP using the same criteria that were used to validate the original programs. In Section 3, we presented the results of a norm-referenced analysis using RMC's original criteria for effectiveness. We did not, however, give the same tests as were used for validation in the original sites. Thus, we felt it important to compare the MAT with those tests. Table 5-7 shows the validating tests used, by program and grade. We wished to determine whether the validating tests were aligned more closely than was the MAT with the PIP-specified curriculum. We could then determine whether differences in the way they aligned with the PIP curriculum would account for the difference in test gains.

The skills tested must be discussed at a more general level than was the congruence elaborated in Appendix G, because a fine-grained analysis was not conducted. We have no records documenting what curriculum was used at originating projects, only what was specified by the PIPs. Consequently, our argument will turn on other judgments as well as extension from the analyses given previously.

To compare the validating tests with the PIP curriculum, we used the general level skills on the MAT as reference points. For the Primer MAT, for example, we isolated the following five skills: (1) matching beginning sounds with pictures, (2) matching ending sounds with pictures, (3) matching beginning sounds with letters, (4) matching ending sounds with letters, and (5) matching spoken words with written words. Skills of a similar quality were developed for the validating tests. Whenever

Table 5-7

TESTS USED TO VALIDATE EFFECTIVENESS OF ORIGINAL PROGRAMS

PIP	Validating Test
Catch-Up	Metropolitan Achievement Test (grades 1-3) Comprehensive Test of Basic Skills (grades 4-6)
Conquest	Gates-MacGinitie (grades 1-3) California Achievement Test, 1957 (grades 4-6)
HIT	Wide Range Achievement Test (grades 6-8)
IRIT	California Achievement Test, 1970 (grade 3)
PTR	Gates-MacGinitie (grade 1)
R-3	Comprehensive Test of Basic Skills (grade 8)

possible, we used the same list of skills for the validating tests as we had used for the MAT; however, when some items on the validating test could not be described by these skills, we added the necessary new categories of skills.

When all skills were defined, it became apparent that many of the skills listed for the MAT were almost the same as those listed for the validating tests. The skills were then reviewed to see which ones, if learned by a student, would allow him to answer more items correctly. The percentage of items devoted to each skill was computed for both Total Reading and Total Math. For each test, this computation was made by dividing the number of items in each reading or math skill category by the number of items devoted to reading or math skills in all of the subtests.

Finally, the skills were reviewed for coverage in the PIP curriculum. We were somewhat limited in this analysis because we had not examined all of the specified materials, only the specified materials that were used at the field-test sites. However, we were generally familiar with the nature of the specified materials and the skills they covered. Although we lacked conclusive evidence, we were confident that some of the skills covered by some of the tests were not covered by the PIP curriculum materials. These skills were marked as a "no" or a "not certain," depending on how confident we were that they were not covered.

Table 5-8

COMPARISON OF POSTTEST TOTAL READING CONTENT BETWEEN THE MAT USED IN THE PIP EVALUATION AND TESTS USED IN EVALUATION OF ORIGINATING PROGRAMS

a. Grades 1 and 2

Test/Content	Grade 1						Grade 2			
	MAT Primer Form F (T = 72) PTR (Canton only)		MAT Primary I Form F (T = 77) PTR (Dallas only) Catch-Up,† Conquest		Gates-MacGinitie* Form 1, Primary A (T = 82) Conquest, PTR		MAT Primary II Form F (T = 84) Catch-Up,† Conquest		Gates-MacGinitie* Form 1, Primary B (T = 82) Conquest	
	Percent of Total Test	PIP-Specified	Percent of Total Test	PIP-Specified	Percent of Total Test	PIP-Specified	Percent of Total Test	PIP-Specified	Percent of Total Test	PIP-Specified
Total Reading	100%		100%		100%		100%		100%	
Listening for Sounds	54									
Match beginning sound with picture	15%	Yes								
Match ending sound with picture	15	Yes								
Match beginning sound with letter	7	Yes								
Match ending sound with letter	4	Yes								
Match spoken with written word	13	Yes								
Word Knowledge			45		59		48		59	
Match word with picture			45%	Yes	59%	Yes	20%	Yes	59%	Yes
Match written word with written word							28	Yes		
Definition							(23)			
Opposites							(5)			
Reading	46		55		41		52		41	
Match sentence with picture	7	Yes	17	Yes	27	Yes	15	Yes	5	Yes
Match story (2-5 sentences) with picture					14	Yes			36	Yes
Single word answer to riddle			11	NC						
Single paragraph stories with questions			27	Yes			37	Yes		
Literal			(15)				(25)			
Inferential			(12)				(8)			
Main idea							(4)			
Recognition of letter names	15	Yes								
Match word with picture	24	Yes								

Note: The Gates-MacGinitie does not collapse vocabulary and comprehension into a Total Reading score at the Primary A level. Nevertheless, percents for Total Reading are displayed because the Dissemination and Review Panel does not state which scores were used to determine the exemplary program.

Key: T = number of test items; NC = not certain. Examination of fourth and eighth grade curricula in sample classrooms provided some anchors from which skills at other grade levels could be extrapolated. Since a thorough search was not conducted at other grade levels, the notation "NC" is used to indicate those skills for which there was no evidence.

*Test used in evaluation of originating program.

†MAT also used for evaluation of originating program.

Table 5-8 (Continued)

5. Grades 3 and 4

Test Content	Grade 3: MAT Elementary Form F (T = 95) Catch-Up,[1] Conquest, IRIT — Percent of Total Test	PIP-Specified	Gates-MacGinitie* Form 1, Primary C (T = 100) Conquest — Percent of Total Test	PIP-Specified	California Achievement Test* Form A, Level 2, 1970 (T = 88), IRIT — Percent of Total Test	PIP-Specified
Total Reading	100%		100%		100%	
Word Knowledge	53		52		47	
Match word with picture			12%	Yes		
Match written word with written word		Yes	40	Yes	24%	Yes
Definition						
Match word in phrase with synonym						
General vocabulary						
Specialized words						
Mathematics						
Science						
Social studies						
Opposites	4	Yes			23	Yes
Match spoken word with written word						
Reading Comprehension	47		48		53	
Passages with questions						
Literal	17	Yes	27	Yes	23	Yes
Inferential	[illegible]	Yes	19	Yes	11	Yes
Main idea	[illegible]	Yes			2	Yes
Word in context	[illegible]	No				
Sequence of events						
Following directions						
Reference skills						
Organization of topics			2	NC	5	NC
Miscellaneous						
Alphabetizing, index, and					6	No
table of contents					6	No
Poetry						
Passage features of stories						
Readability (in grade level)			Passages too short for readability analysis			
[illegible]	[illegible]	Yes				
[illegible]	[illegible]	Yes			19	Yes
[illegible]	[illegible]	Yes				
[illegible]	[illegible]	Yes			11	Yes
[illegible]	[illegible]	NC			10	NC
[illegible]	[illegible]	NC				
[illegible]	[illegible]	NC				
[illegible]	7	NC				
[illegible]						
Mean sentence length (in words)						
[illegible]	[illegible]	Yes			4	Yes
[illegible]		Yes			16	Yes
[illegible]	[illegible]	Yes				
[illegible]	[illegible]	Yes				
[illegible]	[illegible]	Yes				
[illegible]	[illegible]	Yes	32	Yes		
[illegible]	[illegible]	Yes				
[illegible]	[illegible]	Yes	16	Yes	20	Yes
[illegible]						
Passage length (in words)			Length in most passages, 55 ± 10.			
[illegible]	[illegible]	[illegible]			3	Yes
[illegible]	[illegible]	[illegible]			5	Yes
[illegible]	[illegible]	[illegible]				
[illegible]		[illegible]			21	Yes
[illegible]	[illegible]	[illegible]			9	Yes

[illegible]

[illegible]

Table 5-8 (Continued)

b. Grades 3 and 4 (Concluded)

Test/Content	Grade 4: MAT Elementary Form F (T = 95) Catch-Up, Conquest, IRIT† — Percent of Total Test	Grade 4: MAT Elementary Form F (T = 95) Catch-Up, Conquest, IRIT† — PIP-Specified	Grade 4: California Achievement Test* (4,5,6--1957) (T = 120) Conquest — Percent of Total Test	Grade 4: California Achievement Test* (4,5,6--1957) (T = 120) Conquest — PIP-Specified	Grade 4: California Test of Basic Skills* Form Q, Level 2 (4,5,6) (T = 85) Catch-Up‡ — Percent of Total Test	Grade 4: California Test of Basic Skills* Form Q, Level 2 (4,5,6) (T = 85) Catch-Up‡ — PIP-Specified
Total Reading	60%		100%		100%	
Word Knowledge	53		42		47	
Match word with picture						
Match written word with written word						
Definition	49%	Yes				
Match word in phrase with synonym					47%	Yes
General vocabulary			11%	Yes		
Specialized words						
Mathematics			10	NC		
Science			11	No		
Social studies			10	No		
Opposites	4	Yes				
Match spoken word with written word						
Reading Comprehension	47		58		53	
Stories with questions						
Literal	17	Yes	9	Yes	18	Yes
Inferential	22	Yes	6	Yes	11	Yes
Main idea	4	Yes	2	Yes	6	Yes
Word in context	4	No			1	No
Sequence of events			3	Yes		
Following direction			17	Yes		
Reference skills			17	No		
Organization of topics			4	No		
Miscellaneous					4	NC
Alphabetizing, index, and table of contents						
Poetry					13	No
Passage features of stories						
Vocabulary range (by grade level)						
-2.5	7	Yes				
2.6-3.2	5	Yes				
3.3-3.9	4	Yes				
4.0-4.2	4	Yes				
4.3-4.6	8	Yes				
4.7-5.6	5	NC			17	NC
5.7-5.8	6	NC	7	NC		
[illegible]	7	NC			23	NC
6.[illegible]+			10	NC		
Mean sentence length (in words)						
8.[illegible]-9.2	7	Yes				
9.3-[illegible]	7	Yes				
10.7-10.8	5	Yes				
10.9-11.0	6	Yes				
11.1-11.2	6	Yes				
11.3-13.1	5	Yes			10	Yes
13.2-13.6	7	Yes				
13.7-14.8	6	Yes	13	Yes		
14.9+			4	NC	30	NC
Passage length (in words)						
55-65	22	Yes				
66-77	6	Yes				
78-7[illegible]	6	Yes				
80-82	7	Yes				
83-8[illegible]	6	Yes			6	Yes
90+			17	Yes	34	Yes

*Test used in evaluation of originating program.

†No comparison is available for fourth grade IRIT PIP — only third graders were tested at the originating site.

‡The range of vocabulary levels rather than the absolute level of each passage is displayed for this comparison. Vocabulary ranges in the CTBS were framed by the ranges in the MAT passages. While the MAT ranges change from the fourth grade (elementary) to the fifth and sixth grades (intermediate), the same CTBS was administered to all three grades at the originating site. The differences in the CTBS ranges on the tables for the fourth and the fifth-sixth grades represent an accommodation to the changes in the MAT ranges; the absolute vocabulary levels in the CTBS passages remain constant across grades.

Table 5-8 (Continued)

c. Grades 5 and 6

Test/Content	Grades 5 and 6: MAT Intermediate (T = 95) Catch-Up, Conquest, HIT[†] — Percent of Total Test	PIP-Specified	California Achievement Test* (4,5,6--1957) (T = 120) Conquest — Percent of Total Test	PIP-Specified	California Test of Basic Skills* Form Q, Level 2 (4,5,6) (T = 85) Catch-Up — Percent of Total Test	PIP-Specified	Grade 6 Only: Wide Range Achievement Test* Level 1 (6,7,8) (T = 75) HIT — Percent of Total Test	PIP-Specified
Total Reading	100%		100%		100%		100%	
Word Knowledge	53		42		47		100	
Match written word with written word								
Definition	49%	Yes						
Match word in phrase with synonym					47%	Yes		
General vocabulary			11%	Yes				
Specialized vocabulary								
Mathematics			10	NC				
Science			11	No				
Social studies			10	No				
Opposites	4	Yes						
Read list of words aloud							100%	Yes
Reading comprehension	47		58		53		0	
Stories/questions								
Literal	8	Yes	9	Yes	16	Yes		
Inferential	16	Yes	6	Yes	11	Yes		
Main idea	[illegible]	Yes	2	Yes	6	Yes		
Word in context	13	No			1	No		
Sequence of events			3	Yes				
Following directions			17	NC				
Reference skills			17	No				
Organization of topics			4	No				
Miscellaneous					4	NC		
Poetry					13	No		
Passage features of stories								
Vocabulary range (by grade level)								
- 5.2	4	Yes			16	Yes		
5.3- 5.3	6	Yes	7	Yes				
6.4- 6.6	7	NC			23	NC		
6.7- 6.8	7	NC						
6.9- 7.[illegible]	[illegible]	NC						
7.1- [illegible]	[illegible]	NC						
[illegible]	[illegible]	NC						
[illegible]-10.[illegible]	5	No	10	No				
Mean sentence length (in words)								
11.[illegible]-13.0	4	Yes			10	Yes		
13.1-15.2	[illegible]	Yes	13	Yes				
15.3-16.[illegible]	4	Yes			8	Yes		
16.1-17.[illegible]	14	Yes						
17.1-18.[illegible]	5	Yes	4	Yes	9	Yes		
18.8-19.[illegible]	[illegible]	Yes			7	Yes		
19.1+					6	Yes		
Passage length (in words)								
61-[illegible]	[illegible]	Yes			6	Yes		
[illegible]	[illegible]	Yes						
101-1[illegible]	[illegible]	Yes						
113-1[illegible]	[illegible]	Yes						
151-1[illegible]	[illegible]	Yes	13	Yes				
176-1[illegible]	[illegible]	Yes			8	Yes		
190-20[illegible]	[illegible]	Yes						
202+			[illegible]	Yes	[illegible]	Yes		

*Test [illegible]

[†]HIT, [illegible]

Table 5-8 (Concluded)

d. Grades 7 and 8

Test/Content	MAT Advanced Form F (T = 95) HIT, R-3† — Grades 7 and 8: Percent of Total Test	PIP-Specified	Wide Range Achievement Test* Level 1 (6,7,8) (T = 75) HIT — Grades 7 and 8: Percent of Total Test	PIP-Specified	California Test of Basic Skills* Form Q, Level 3 (T = 85) R-3 — Grade 8 Only: Percent of Total Test	PIP-Specified
Total Reading	100%		100%		100%	
Word Knowledge	53		100		47	
Match written word with written word						
Definition: match word in phrase with synonym	49%	Yes			47%	Yes
Opposites	4	Yes				
Read list of words aloud			100%	Yes		
Reading Comprehension	47		0		53	
Stories/questions						
Literal	7	Yes			8	Yes
Inferential	23	Yes			17	Yes
Main idea	4	Yes			7	Yes
Word in context	13	No			2	No
Miscellaneous					6	NC
Understanding rules and poetry					13	NC
Passage features† t stories						
Vocabulary range (by grade level)						
- 7.6	5	Yes			17	Yes
7.7- 8.1	6	Yes				
8.2- 8.8	4	NC			13	NC
8.9- 9.1	9	NC			9	NC
9.2-10.5	9	NC				
10.6-10.7	7	NC				
10.8-11.5	7	NC				
Mean sentence length (in words)						
12.2-12.6	7	Yes				
12.7-14.4	6	Yes			8	Yes
14.5-14.8	8.5	Yes				
14.9-15.0	9.5	Yes				
15.1-16.3	16	Yes			31	Yes
Passage length (in words)						
105-149	4	Yes			18	Yes
150-159	5	Yes				
160-202	8	Yes			21	Yes
203-261	6	Yes				
262-303	24	Yes				

*Test used in evaluation of originating program.

†See [illegible]

Table 5-9

COMPARISON OF POSTTEST TOTAL MATHEMATICS CONTENT BETWEEN THE MAT USED IN THE PIP EVALUATION AND TESTS USED IN EVALUATION OF ORIGINATING PROGRAMS

	Catch-Up, Grade 4		Catch-Up, Grades 4,5,6		Catch-Up, Grades 5,6; HIT, Grade 6		HIT, Grades 6,7,8		HIT, Grades 7,8; R-3, Grade 8		R-3, Grade 8	
	MAT Elementary Form F (T = 115)		California Test of Basic Skills* Form Q, Level 2 (T = 98)		MAT Intermediate Form F (T = 115)		Wide Range Achievement Test* Level 1 (T = 43)		MAT Advanced Form F (T = 115)		California Test of Basic Skills* Form Q, Level 3 (T = 98)	
Test Content	Percent of Total Test	PIP-Specified	Percent of Total Test	PIP-Specified	Percent of Total Test	PIP-Specified	Percent of Total Test	PIP-Specified	Percent of Total Test	PIP-Specified	Percent of Total Test	PIP-Specified
Total Math	100%		100%		100%		100%		100%		100%	
Computation	35		50		36		100		35		49	
Basic operations: whole numbers	29%	Yes	27%	Yes	17%	Yes	33%	Yes	12%	Yes	8%	Yes
Fractions, decimals only	6	Yes	23	Yes								
Fractions, decimals, percents					17	Yes	42	Yes	16	Yes	38	Yes
Measurement							9	Yes				
Other: rounding, averages, root, exponents, equations, negative numbers					2	No	16	No	7	†	3	ND
Concepts	35		30		34		0		35		31	
Traditional math. operations, measurement, place value	16	Yes	20	Yes	11	Yes			9	Yes	16	Yes
Modern math: sets, equations, estimations, array, laws, number series, geometry, notation	19	No	10	No	23	No			26	†	15	ND
Problem Solving	30 30	NC	20 20	NC	30 30	NC	0		30 30	†	20 20	ND

Key: T = number of items in original validation sample.

ND = not determined for R-3; R-3 math allows much flexibility for teachers and, unlike other PIPS, is not supplementary to the regular classroom. Because the content of the curriculum is largely unknown, the notation "ND" is used.

NC = not certain; examination of fourth and eighth grade curricula in sample classrooms provided some anchors from which skills at other grade levels could be extrapolated. Since a thorough search was not conducted at other grade levels, the notation "NC" is used to indicate those skills for which there was no evidence.

*Test used in evaluation of originating program at indicated grade levels.

† = No for HIT; ND for R-3.

The content of the validating tests is compared with the content of the MAT in Table 5-8 (for posttest Total Reading) and in Table 5-9 (for posttest Total Math). These tables show the skills covered by each test and the percentage of items devoted to each skill. Skills are grouped according to the subtests in the MAT. When added, the "percent of total test" for all skills within a subtest will equal the "percent of total test" for that subtest. The total number of reading items (or of math items) is listed as "T" at the top of each column. In the following sections, we discuss the results of our analysis of Tables 5-8 and 5-9.

Catch-Up--Catch-Up validation was based on the same levels of the MAT for grades 1-3 as were used in our analysis, but the validation for grades 4-6 was based on the California Test of Basic Skills (CTBS). Because the tests for the first three grades were the same, they were equally appropriate and the test scores should be comparable.

The CTBS and the MAT have a somewhat different format for word knowledge items. The CTBS tends to be a little easier because it gives the target word in a phrase that helps clarify the meaning. This format also matches the PIP curriculum better than does the MAT because most materials required the student to determine the meaning from the passage.*

The format for the CTBS items on reading comprehension is similar to that of the MAT except that the CTBS includes a section on poetry and the MAT has twice as many inferential questions. Grades 4-6 at the originating site all took the same level of the CTBS, whereas the students at the field-test projects switched from the Elementary MAT to the Intermediate MAT in the fifth grade. At Grade 4, the level of the MAT passages is closer to the fourth grade materials than the CTBS, which is substantially more complex. In grades 5 and 6 the positions are reversed, and the CTBS passages more closely match the curriculum. The confounding factor for grades 5 and 6 is the poetry section on the CTBS, and for grade 4 is the larger number of inferential questions. Even so, this does not outweigh the differences in the passage levels.

The CTBS Total Math subtest emphasizes computation more than does either the Elementary or the Intermediate MAT. The CTBS and the Intermediate MAT emphasize whole numbers and fractions; the Elementary MAT emphasizes whole number operations. As stated earlier, the Catch-Up curriculum focused most heavily on computation skills. It would seem that

---

*However, students are rarely tested on this ability in the exercises that we examined for the fourth and eighth grades.

the CTBS is more relevant than either MAT battery, but in grade 4 (Elementary MAT), where most students were working with whole numbers, the extra items on fractions in the CTBS could mean that it is the less relevant test.

The heavy emphasis in the MAT on both concepts and problem solving was not reflected in Catch-Up curriculum. The CTBS seems to be closer to the PIP curriculum because it does not emphasize either subject, and the items that do deal with concepts and problem solving use the traditional math style.

Overall, the CTBS seems to be more appropriate to Catch-Up's PIP-specified curriculum. Assuming that this similarity would help students to answer more of the test items correctly, we feel that (except for grade 4) the project might have proved more effective in terms of both reading and math scores if the validating test had been used.

Conquest--The validation for Conquest was based on three levels of the Gates-MacGinitie for grades 1-3. The replicating sites were tested on three levels of the MAT.

The Gates-MacGinitie has a larger percent of word knowledge items than do any of the MAT batteries. The Primary II and Elementary (grades 2 and 3) batteries of the MAT emphasize matching written words with written words, while the comparable levels of the Gates-MacGinitie continue to use some items that simply match words with pictures. The matching of a picture with a word was a little more common in the Conquest curriculum than the matching of written words. Thus, for grades 2 and 3 we believe the Gates-MacGinitie is more closely aligned with the PIP-specified and used curriculum. The word knowledge sections of both tests seem to be equally appropriate for grade 1.

A similar comparison about pictures can be made relative to the reading comprehension items for grades 1 and 2. Some Gates-MacGinitie items simply require matching a story of two to five sentences with a picture, while the MAT items compel the student to read a single paragraph story and answer some questions about it. By grade 3 both tests have stories with questions, but the Gates-MacGinitie has only literal and inferential questions, whereas the MAT also has main-idea and word-in-context items. The Conquest curriculum contained a fair amount of reading comprehension materials with a picture format, which brings it closer to the Gates-MacGinitie in grades 1 and 2. In grade 3 the added types of comprehension questions on the MAT again makes the Gates-MacGinitie the more PIP-relevant test.

Overall, in grades 1-3 the Gates-MacGinitie appears to be the better test for the Conquest reading curriculum.

All students in grades 4-6 at the originating site were tested on the same level of the 1957 California Achievement Test (CAT). At the replicating projects, we administered the Elementary MAT to grade 4 and the Intermediate MAT to grades 5 and 6. The difference in the dates of the two tests, the 1957 CAT and the 1970 MAT, will influence the results of this comparison because the older test reflects a somewhat different content emphasis in testing.

The MAT stresses word knowledge skills more than does the CAT, but the CAT includes more specialized words from mathematics, science, and social studies. Conquest did not emphasize any specialized words, and so the MAT seems to be the more appropriate test.

Although a larger portion of Total Reading content is devoted to reading in the CAT, only 20% of the test deals with reading stories and answering comprehension questions. About 38% of the Total Reading items cover skills like following directions, reference skills, and organization of a topic; although some Conquest materials covered following directions, we found none that dealt with the other two skills.

Overall, even though we found Conquest materials that followed the reading comprehension format of both the MAT and the CAT, we feel the CAT to be the less relevant test for reading skills because it includes skills not covered in the curriculum. For this reason, students should have scored higher on the MAT.

HIT--HIT was evaluated by the same level of the Wide Range Achievement Test (WRAT) for grades 6-8. The Intermediate MAT was used to test grade 6 in our analysis; the Advanced MAT was used for grades 7 and 8.

All reading items on the WRAT deal with reading a list of words aloud. The test has no word knowledge questions and no stories with comprehension questions. The HIT curriculum emphasized phonics and oral drill work. The curriculum offered little reading comprehension material and did not cover most of the vocabulary in the Advanced MAT Word Knowledge subtest. The WRAT is much more closely aligned with the emphasis in the HIT reading curriculum.

All math items on the WRAT are computation problems. Of the basic operation items, 42% entail working with fractions, and 33% with whole numbers; the MAT places less emphasis on fractions. Because the WRAT excludes math concepts and problem solving and because the HIT curriculum emphasized computations using both whole numbers and fractions but did not emphasize either math concepts or problem solving, we believe the WRAT more closely parallels the focus of the HIT project.

In addition, the WRAT is substantially shorter than the MAT batteries. The WRAT has 118 items across both reading and math and takes 30 minutes to administer. The MAT has 210 items and takes 125 minutes to administer. These are the only two tests for which we feel that the difference in the number of items and the testing time would affect test scores. The MAT is substantially more difficult and more taxing of students' abilities than is the WRAT.

Overall, the WRAT is more appropriate for HIT curriculum. With the added feature of its short length, we believe that the field-test project would have seemed more effective if HIT students had been tested on the WRAT.

IRIT--IRIT was originally validated on the 1970 CAT for only grade 3. The third grade students in our study were given the Elementary MAT. We find it difficult to say what skills the IRIT project was trying to emphasize because it specified so many materials. We do know that IRIT's curriculum was divided into three sections: phonics, vocabulary and comprehension, and individualized reading. Phonics is not covered by either test, but the other two areas are covered by both.

Although the MAT gives the word knowledge skills a larger portion of the Total Reading score, the CAT has two formats for these items. Half of the items require matching a spoken word with a written word; the other half, which uses the same format as the MAT, requires matching a written word with a similar written word. We are unable to say if one of these formats received more emphasis in the IRIT curriculum. Relative to word knowledge, both tests appear to be equally appropriate for the IRIT curriculum.

The MAT features more items on reading comprehension than does the CAT. The MAT's emphasis is on inferential questions, the CAT's on literal questions. The IRIT materials revealed a slight preference for literal questions. CAT passage features not only are lower than MAT features, but also are probably at a level more commonly encountered by third grade students. As does its earlier edition, the 1970 CAT tests some skills that do not relate as directly to reading and are not obviously covered in the IRIT curriculum. The CAT has items that are missing on the MAT, such as reading a table of contents and an index. Even though the passages on the MAT are more difficult than the ones in the IRIT materials, the percentage of CAT items dealing with skills not covered by IRIT appears to make the CAT the less relevant reading comprehension test.

Overall, however, the CAT is an easier test so, even though the CAT has some skills not covered in the IRIT curriculum, students at the field-test sites might have done better if they had taken the CAT.

PTR--The Gates-MacGinitie was used to validate the originating PTR site. In our evaluation, we tested grade 1 in Canton with the MAT Primer and grade 1 in Dallas with the Primary I.

None of the skills in the Listening for Sounds subtest in the Primer are covered on the Gates-MacGinitie, but they were extensively covered in the PTR curriculum materials.

The word knowledge items in the Primer were included in the Reading subtest. The Gates-MacGinitie puts a heavier emphasis on word knowledge than either of the MAT batteries. The PTR curriculum included lessons on word knowledge that are similar to items on both tests.

The composition of the Reading subtest is different for all three tests (Gates-MacGinitie, Primer, and Primary I). The only item common to all is that of matching a sentence with a picture. The other items on the Primer examine recognition of letter names. The other two sections of the Primary I deal with riddles and with reading a single paragraph and answering comprehension questions. The other section of the Gates-MacGinite requires matching a picture with a story of two or three sentences. The PTR curriculum did not cover riddles, and the single paragraph with questions and the matching of a picture with a multiple-sentence story are found only at the more advanced levels. The curriculum placed tremendous emphasis on recognition of letter names. The single-sentence picture match was also covered in the curriculum, but not as heavily. Of the three tests, the Primer is probably the best reading test for PTR because it emphasizes the same skills as the curriculum. The Gates-MacGinitie would be the next best test, and the Primary I the least appropriate.

Overall, the Primer appears to be the best test of the PTR program because it covers the skills emphasized in the curriculum and includes more of the skills on which PTR focused. The Gates-MacGinitie is less appropriate, and the Primary I, the least appropriate.

R-3--The test for validating the original R-3 project was the CTBS. For our evaluation, R-3 students were given the Advanced MAT.

Word knowledge skills are emphasized equally on the MAT and the CTBS, but the format on the CTBS is different. This format, which gives the target word in a phrase that helps clarify the meaning, more closely matches the curriculum materials.

The reading comprehension items receive approximately equal emphasis on both tests, with CTBS having additional sections on miscellaneous skills and on understanding rules and poetry. The MAT passages are more difficult than the ones in the CTBS, and the MAT places greater emphasis on word-in-context and inferential items. We were unable to find any evidence that the R-3 materials covered skills like understanding rules or poetry. The more difficult passages in the MAT were not reflected in the curriculum, with the possible exception of the most advanced levels. Although we could not determine from the curriculum how the MAT emphasis on inferential questions would have affected the students' scores, the lack of word-in-context skills in the curriculum could have had a negative effect.

The CTBS stresses math computation more than does the MAT, especially basic operations using fractions. Although both tests emphasize math concepts, the MAT has many more items on modern math. The MAT also has a much heavier emphasis on problem solving. However, the R-3 curriculum covered such a tremendous variety of skills that it is difficult to say what skills would be considered core to this program. Because of this variety in math skills and formats, one test cannot be designated the more appropriate for this program.

Overall, we feel that R-3 students could have done equally well on both tests. For example, in the Reading subtest the increased difficulty of the MAT stories is balanced by the additional skills required on the CTBS. Relative to math, lack of knowledge about what was core to the program prevents distinguishing between tests.

5.9 Conclusions

In 10 of the 19 PIP/grade combinations, the validating test provided a closer match with curriculum and a better chance of showing project effectiveness. In five PIP/grade combinations, the MAT was the more appropriate test. In the remaining four combinations, neither the pretest nor the posttest could be rated substantially better; three of these combinations are for Catch-Up, grades 1-3, where the MAT was used for both pre- and post-testing.

In our analyses, we attempted to verify that there was reason to believe that successful projects would increase MAT scores. Such verification was dependent on showing that the MAT was congruent with the curriculum specified by the PIP and used with students.

From a rather gross analysis of MAT skills, the MAT appeared relevant to PIP curricula, even though PIP projects exhibited considerable diversity in their selection of teaching materials (e.g., IRIT projects, which showed almost no overlap in materials). The diversity of curricula may not be counter to RMC's expectations, since they packaged the "programs" of the originating sites even when such programs had no consistent instructional methods or specific curriculum materials. However, we regard this diversity as an unsatisfactory outcome for prescriptive packages that promised to cause the same achievement effects. At the same level of analysis, we found that (except for R-3) the PIP project curricula were not essentially different from each other or from the regular school curricula. That is, the PIPs did not transport fundamentally innovative projects, although they did establish working projects. This is not a limitation of packaging, but rather a consequence of what was selected for packaging.

From our fairly detailed analysis of the relationship of the MAT to fourth and eighth grade curricula and our less detailed analysis of the relevance of the validating tests to all curricula observed, we concluded that for most PIP projects the validating test would have been more responsive to the curricula. Our analysis of fourth and eighth grade curricula showed that only the MAT Math Computation subtest seemed particularly relevant to PIP objectives. In our norm-referenced results (Section 3), the math scores stand out as an area of project success compared with MAT Total Reading; this confirms that some of what was taught was learned and, given a relevant test, the evaluator can detect that learning had occurred. We are not surprised that scores were low in the areas not covered by the PIP because, according to our limited information, these areas were not covered by regular classroom curricula either.

We have found no reason to believe that MAT scores should be greatly increased by participation in PIP projects. PIP curricula are not innovative, but were supportive of the regular curriculum, so that credit for the limited successes of the PIP projects must be shared with the regular classes.

The main variables manipulated by PIP projects to improve MAT scores over scores expected from regular instruction appear to have been classroom management and lower student/teacher ratios. PIP-induced curricula did not teach the child anything relevant to the MAT that he could not learn from other sources. Most important, the MAT was not found to be particularly relevant for assessing the achievement impact of PIPs.

6 THE EFFECT OF IMPLEMENTATION ON ACHIEVEMENT

6.1 Introduction

Because the MAT was not particularly relevant to the curricula of PIP projects, we had no specific reason to expect large gains in MAT scores. Nevertheless, because these scores were the only measure we had for PIP effects on student achievement, we present some formal analyses relating them to teacher implementation and responsiveness (described in Section 4). Preparatory to these analyses, we discuss some simulations done to guide our choice for the metric of the dependent variable. We conclude that the MAT standard score metric is defective and is therefore inappropriate for evaluative purposes.

Our main conclusion is that the formal analyses reported below do not support the claim that PIP implementation alone produces large gains on our measurement of achievement. Teacher responsiveness is more often effective.

In this evaluation, generally as in others of its scope, no sampling of project teachers, students, or locations was possible. As a consequence, the empirical justification for the usual inferential statistical techniques is not present. The basis for our analyses is not that of inferential statistics, but that of curve fitting and "least squares" descriptive statistics. Consequently, we will not report our results according to the canons of statistical decision theory. The problem with the inferential framework in this study is that with no sampling scheme we have no basis for claiming that the probability statements associated with hypothesis tests have any empirical significance.*

6.2 Definition of Regression Model for Teacher Implementation and Responsiveness Ratings

In Section 4 we described how we rated teachers on two factors: implementation of project and responsiveness. These variables were nominally scored according to the scheme shown in Table 6-1. The ratings were then converted to three "dummy variables" per project, as follows:

---

* We do not adopt a Bayesian point of view for the corresponding reason: We have no satisfactory posterior distributions.

Rating	Well Implemented (I_W)	Good Responsiveness (I_G)	Well Implemented or Good Responsiveness ($I_{W/G}$)
1	1	1	1
2	1	0	1
3	1	0	1
4	0	1	1
5	0	0	0
6	0	0	0
7	0	1	1
8	0	0	0
9	0	0	0

Table 6-1

CLASSIFICATION SCHEME FOR PROJECT TEACHERS DURING OBSERVATION

Teacher's Implementation of Project	Teacher's Responsiveness		
	Good	So-So	Bad
Well implemented	1	2	3
So-so implemented	4	5	6
Poorly implemented	7	8	9

Most of the analyses reported below entail only I_W and I_G; when this caused singularities, $I_{W/G}$ was used. The distribution of teachers on these variables may be inferred from Table 4-5 in Section 4.

Our basic descriptive model is a bivariate regression equation:

$$Y_{ijkm} = B_{0k} + B_{1km} I_{W_{jm}} + B_{2km} I_{G_{jm}} + B_{3k} I_{S_{ijm}} + B_{4k} I_{R_{ijm}} + B_{5k} A_{ijlm} + B_{ijm} I_{ijm} + \varepsilon_{ijkm} , \qquad (6\text{-}1)$$

where

$m = 1, M$ Indexes projects.

$k = 1, 2$ Indexes fall and spring observations, respectively.

$j = 1, J_m$ Indexes teachers within projects, when there are J_m teachers.

$i = 1, T_{jm}$ Indexes students of the jm^{th} teacher, when there are T_{jm} students for the jm^{th} teacher.

$$I_{W_{jm}} = \begin{cases} 1 & \text{If the } jm^{th} \text{ teacher had a well-implemented project during our site visits.} \\ 0 & \text{Otherwise.} \end{cases}$$

$$I_{G_{jm}} = \begin{cases} 1 & \text{If the } jm^{th} \text{ teacher was responsive during our site visits.} \\ 0 & \text{Otherwise.} \end{cases}$$

$$I_{S_{ijm}} = \begin{cases} 1 & \text{If the } ijm^{th} \text{ student was male.} \\ 0 & \text{Otherwise.} \end{cases}$$

$I_{R_{ijm}}$ Is an indicator variable for the student's race. The exact specification of this variable depends on the ethnic distribution of each project. See Appendix D, which shows the independent variables used in equations for the regression analysis.

A_{ijlm} Is the ijm^{th} student's age in the fall.

I_{ijm} Is an indicator variable for each student. The parameter associated with this variable is shown for completeness sake; we will not estimate it.

ε_{ijkm} Represents the error for student ijm at time k.

Our evidence for PIP effectiveness would be that

$$B_{12m} - B_{11m} > 0 \qquad (6\text{-}2)$$

for a large percentage of our PIP and grade combinations. When this inequality holds, our model asserts that, given the responsiveness of the teachers and the values of the other variables at a PIP and grade, being a student of a teacher with a better implemented project makes a greater impact in the spring than in the fall. If our inequality does not hold, we have no positive association between degree of implementation and outcomes and no evidence that implementing the PIP well is associated with increased values of the dependent variables.

The model (Eq. 6-1) suffers from the defects that the norm-referenced procedure (Section 2.3.2) attempted unsuccessfully to overcome. The model does not tell us what would occur if no project were in place, nor does it say how far from zero the inequality (Eq. 6-2) must be to be educationally significant. A problem neither expression (Eq. 6-1 nor Eq. 6-2) addresses is how many successes according to the model would imply that the PIPs are successful.

6.3 Selection of Metric for the Outcome Variable

As a result of the content analyses presented in Section 5, we know that the MAT items are not highly relevant to the PIP curricula, and, as a result of the analyses described in Section 2, we know that the longitudinal validity of the MAT norms is questionable. The implication is that the MAT standard scores may not be an appropriate metric for the analysis of PIP project achievement outcomes.

The "grade effects" on the Reading subtest discussed in Section 3.7.1 can be interpreted as further evidence of this. In that section we found that the percentile of first and fourth grade Reading subtest averages declined as a function of time over a wide variety of project types, locations, and student body characteristics. If these declines are artifacts (i.e., if the declines do not reflect some defect common to all projects at these grades), obviously the MAT standard score metric should be abandoned.

Consequently, we decided to investigate whether there might not be artifacts in the MAT standard scores that would cause apparent declines. Our investigation was conducted by means of the simulations described in the next section.

6.3.1 Simulation of the Norm-Referenced Analyses for the MAT Reading Subtest--Grades 3, 4, and 5

We had observed in the PIP data a definite trend toward gains in percentile of project averages in the third and fifth grades and losses in the fourth. Obviously, if the MAT norms were valid in 1970, the declines could be the reflection of some developmental factor that came into play in children ten years old in 1976, but that was not present in 1970. Alternatively, if the norms are currently valid, these declines might mean a serious defect in all fourth grade PIP curricula. A third alternative is that the fourth grade MAT norms are not, and never were, valid. Evidence for this view can be obtained from the Anchor Test Study (1974). At the fourth grade the Anchor Study percentiles for reading between the 60th and 2nd (see Table 28 in Anchor Test Study "Equivalence and Norm Tables for Selected Reading Achievement Tests" (1974).

For our simulations, we created raw score distributions in which the "effect" of a program was to increase by a fixed proportion, B, the number of items answered correctly.* The question is, will programs that are equally effective in this sense, be equally effective in the sense of the norm-referenced analysis for reading that we used for the PIP projects, no matter what member of the MAT battery is used.

It is not necessary that programs that are equally effective on a MAT raw score metric be equally effective on the MAT standard score metric if the projects are tested using different members of the MAT battery. However, if they are not, the inequality is evidence that the standard score transformation is not of the same form at each MAT level. If the transformations are not of the same form, the underlying distributions are not comparable. This may mean that the underlying traits being measured are different.

Our simulation generated 3000 pseudorandom variables, P_i, i = 1 to 3000, distributed as the Beta, with parameters α_1 and α_2, so that the mean of the distribution of the simulated fall standard scores, calculated as below, was about what we had observed in the PIP study.

The P_i were converted to Reading subtest standard scores for each grade by the following formula:

$$R_f = [P + G(1 - P)] N_f \quad , \qquad (6\text{-}3)$$

where R_f is the fall raw score, N_f is the number of items on the MAT used in the fall, and G is a guessing parameter.

$$C_s = PN_p + B(N_s - PN_p) \quad , \qquad (6\text{-}4)$$

where N_s is the number of items on the MAT used in the spring, and N_p is the number of items on the spring test that are parallel with items in the fall test. B is the "effect" of the program, as compared with no relevant education at all.

$$R_s = C_s + G(N_s - C_s) \quad , \qquad (6\text{-}5)$$

where R_s is the spring raw score and N_s is the number of items on the spring test.

*These simulations were programmed by George Byrd, Pat McCall, and Roy Sutton of SRI, using IMSL's Beta random number generation.

The 3000 raw score pairs, R_f and R_s, were converted to standard scores through the MAT raw-to-standard score tables as follows:

Grade	Testing	Reading Subtest Table
3	Fall	Primary II
3	Spring	Elementary
4	Fall	Elementary
4	Spring	Elementary
5	Fall	Elementary
5	Spring	Intermediate

The 3000 fall and spring standard scores were used to generate 100 norm-referenced analyses calculated on 30 observations apiece. The same 3000 observations were used at each grade that had the same value of α_1 and α_2. P_i, $30(k - 1) + 1 \leq i \leq 30k$, were used in the k^{th} analysis, $k = 1, 100$.

The results of these simulations are presented in Tables 6-2 and 6-3. Each table shows the percentage of the norm-referenced analyses at each combination of α_1, α_2, and grade, which resulted in the various decisions on the achievement of normal growth and criterion growth. The statistics in the table refer to averages in the norm-referenced analyses.

Table 6-2 shows the results of our simulation for selected values of B, when G = 0. When B = 0, this table shows that the norm-referenced analyses are not subject to grade effects. However, the mean gain over expected is much higher in the fourth grade than in the third and fifth grades. When B = 0.1, the norm-referenced analysis begins to show grade effects, with 51% of the analyses in the fourth grade confirming normal growth, while none confirms normal growth at the other grades. At B = 0.2, 100% of the fourth grade analyses confirm normal growth, while 71% of the third grade and none of the fifth grade analyses show it. Thus, for data like those found in the PIPs, our simulation shows that when G = 0 the tests are differentially sensitive to B, the proportion of items "due to program." In the simulation presented in Table 6-2, one would characterize the fourth grade as easier than the third or the fifth grade, given the B's we have used in our model.

Other simulations confirmed the results when grades were compared using the B's shown on Tables 6-2 and 6-3 with the following α's common to all grades: $\alpha_1 = \alpha_2 = 2.0$; $\alpha_1 = 2.0$, $\alpha_2 = 4.0$: $\alpha_1 = 4.0$, $\alpha_2 = 2.0$.

Table 6-2

PROPORTION OF SIMULATED NORM-REFERENCED ANALYSES FOR MAT READING SUBTEST, WITH THE INDICATED DECISIONS FOR THE ACHIEVEMENT OF CRITERION GROWTH AND NORMAL GROWTH: G = 0.0

		Grade 3* Alpha 1 = 2.0 Alpha 2 = 2.0	Grade 4 Alpha 1 = 2.0 Alpha 2 = 4.0	Grade 5* Alpha 1 = 2.0 Alpha 2 = 2.0
B = 0.00	Meets criterion growth	Yes 0%; No 100%; U 0%	Yes 0%; No 100%; U 0%	Yes 0%; No 100%; U 0%
	Meets normal growth	Yes 0%; No 100%, U 0%	Yes 0%; No 100%; U 0%	Yes 0%; No 100%; U 0%
	Fall score	Mean 45.48; SD 2.17	Mean 48.40; SD 2.94	Mean 60.79; SD 3.00
	Spring score	Mean 31.17; SD 1.67	Mean 48.40; SD 2.94	Mean 43.81; SD 1.62
	Expected score	Mean 46.97; SD 1.98	Mean 53.09; SD 2.68	Mean 66.42; SD 2.86
	Gain over fall	Mean -14.31; SD 0.70	Mean 0.00; SD 0.00	Mean -16.99; SD 1.44
	Gain over expected	Mean -15.80; SD 0.65	Mean - 4.69; SD 0.36	Mean -22.61; SD 1.32
B = 0.10	Meets criterion growth	Yes 0%; No 100%; U 0%	Yes 0%; No 100%; U 0%	Yes 0%; No 100%; U 0%
	Meets normal growth	Yes 0%; No 98%; U 2%	Yes 51%; No 0%; U 49%	Yes 0%; No 100%; U 0%
	Fall score	Mean 45.48; SD 2.17	Mean 48.40; SD 2.94	Mean 60.79; SD 3.00
	Spring score	Mean 41.81; SD 1.78	Mean 54.45; SD 2.25	Mean 55.01; SD 1.18
	Exptected score	Mean 46.97; SD 1.98	Mean 53.09; SD 2.68	Mean 66.42; SD 2.86
	Gain over fall	Mean - 3.67; SD 1.02	Mean 6.06; SD 0.73	Mean - 5.78; SD 1.86
	Gain over expected	Mean - 5.16; SD 0.90	Mean 1.36; SD 0.55	Mean -11.40; SD 1.73
B = 0.20	Meets criterion growth	Yes 1%; No 12%; U 87%	Yes 10%; No 0%; U 90%	Yes 0%; No 78%; U 22%
	Meets normal growth	Yes 71%; No 0%; U 29%	Yes 100%; No 0%; U 0%	Yes 0%; No 16%; U 84%
	Fall score	Mean 45.48; SD 2.17	Mean 48.40; SD 2.94	Mean 60.79; SD 3.00
	Spring score	Mean 50.21; SD 0.77	Mean 59.43; SD 1.75	Mean 64.25; SD 0.85
	Exptected score	Mean 46.97; SD 1.98	Mean 53.09; SD 2.68	Mean 66.42; SD 2.86
	Gain over fall	Mean 4.73; SD 1.43	Mean 11.03; SD 1.25	Mean 3.46; SD 2.18
	Gain over expected	Mean 3.25; SD 1.26	Mean 6.34; SD 1.02	Mean - 2.16; SD 2.04

Note: U = unknown.

*Results in these columns are based on the same 3000 observations distributed as Beta with $\alpha_1 = \alpha_2 = 2.0$.

Table 6-3

PROPORTION OF SIMULATED NORM-REFERENCED ANALYSES FOR MAT READING SUBTEST, WITH THE INDICATED DECISIONS FOR THE ACHIEVEMENT OF CRITERION GROWTH AND NORMAL GROWTH: G = 0.25

		Grade 3 Alpha 1 = 2.0 Alpha 2 = 4.0	Grade 4 Alpha 1 = 0.1 Alpha 2 = 1.0	Grade 5 Alpha 1 = 1.5 Alpha 2 = 5.0
B = 0.00	Meets criterion growth	Yes 0%; No 99%; U 1%	Yes 0%; No 100%; U 0%	Yes 0%; No 100%; U 0%
	Meets normal growth	Yes 99%; No 0%; U 1%	Yes 0%; No 100%; U 0%	Yes 11%; No 0%; U 89%
	Fall score	Mean 47.68; SD 0.88	Mean 49.19; SD 1.90	Mean 56.99; SD 1.26
	Spring score	Mean 50.98; SD 0.46	Mean 49.19; SD 1.90	Mean 63.58; SD 0.52
	Expected score	Mean 48.74; SD 0.86	Mean 53.76; SD 1.77	Mean 62.91; SD 1.13
	Gain over fall	Mean 3.29; SD 0.49	Mean 0.00, SD 0.00	Mean 6.59; SD 0.77
	Gain over expected	Mean 2.23; SD 0.47	Mean - 4.57; SD 0.30	Mean 0.67; SD 0.64
B = 0.10	Meets criterion growth	Yes 97%; No 0%; U 3%	Yes 0%; No 100%; U 0%	Yes 63%; No 0%; U 37%
	Meets normal growth	Yes 100%; No 0%; U 0%	Yes 97%; No 0%; U 3%	Yes 100%; No 0%; U 0%
	Fall score	Mean 47.68; SD 0.88	Mean 49.19; SD 1.90	Mean 56.99; SD 1.26
	Spring score	Mean 55.40; SD 0.37	Mean 55.55; SD 1.56	Mean 69.33; SD 0.33
	Expected score	Mean 48.74; SD 0.86	Mean 53.76; SD 1.77	Mean 62.91; SD 1.13
	Gain over fall	Mean 7.71; SD 0.56	Mean 6.36; SD 0.43	Mean 12.33; SD 0.96
	Gain over expected	Mean 6.65; SD 0.55	Mean 1.80; SD 0.41	Mean 6.41; SD 0.83
B = 0.20	Meets criterion growth	Yes 100%; No 0%; U 0%	Yes 42%; No 0%; U 58%	Yes 100%; No 0%; U 0%
	Meets normal growth	Yes [illegible]; No 0%; U 0%	Yes 100%; No 0%; U 0%	Yes 100%; No 0%; U 0%
	Fall score	Mean 47.68; SD 0.88	Mean 49.19; SD 1.90	Mean 56.99; SD 1.26
	Spring score	Mean 59.37; SD 0.31	Mean 59.33; SD 1.44	Mean 73.53; SD 0.27
	Expected score	Mean 48.74; SD 0.86	Mean 53.76; SD 1.77	Mean 62.91; SD 1.13
	Gain over fall	Mean 11.68; SD 0.61	Mean 10.13; SD 0.53	Mean 16.54; SD 1.01
	Gain over expected	Mean 10.62; SD 0.59	Mean 5.57; SD 0.47	Mean 10.62; SD 0.89

Note: U = unknown.

Table 6-3 shows the results of our simulation for the same values of B discussed above. In this table, G = 0.25, and the α's have been changed from those shown in Table 6-2 to make the fall average scores about the same as were observed in the PIP study.

In Table 6-3, the results for the fourth grade are largely unchanged. The results for the third and fifth grades are dramatically different, however. When we simulated no guessing, the fourth grade test appeared easier in terms of results for a fixed value of B. Table 6-3 indicates that with simulated guessing (G = 0.25) the fourth grade test is now the most difficult. In fact, some analyses confirm normal growth in the third and fifth grades, even when B = 0.

Overall, these grade effects are similar to the grade effects noted in Section 3.7. They indicate that the norm-referenced analysis as applied to MAT reading data gives different results for programs that are equally effective (as measured by B in Eq. 6-4). Which reading test is easier depends on the size of gain that is allowed. If the values we used for R_s are increased in Eq. 6-5 by the second term, the fourth grade is harder than either the third or the fifth. If we do not increase them, the fourth grade is easier. It is important to note that this result, however, is relative because the fourth grade standard scores are nearly stable. It is the third and fifth grades that are sensitive to the changes we have simulated.

These findings show that the items selected for the various MAT batteries were not such that the batteries have comparable distributions of item difficulties in their respective norm groups. Of course, there is no requirement that there be comparable norm group distributions of item difficulties. The Thurstone techniques that are the analytic foundations of the MAT standard scores may be applied no matter what the raw score distributions. Indeed, the standard score transformation may be viewed as correcting for the fact that the norm group distributions differ and that, therefore, equal raw scores on different batteries do not represent equal values on an <u>assumed</u> (unobservable) underlying normally distributed skill continuum.

The usefulness of this correction depends on the reasonableness of the assumption of the existence of the unobservable skill continuum and the reasonableness of the presumed distribution of children's values on this continuum.

We feel that the three reading subtests--that of the Primary II, the Elementary, and the Advanced--are sufficiently dissimilar to justify interpreting the differing distributions of item difficulties as evidence that these subtests are not measuring the same skills. The Primary II Reading subtest entails much lower level reading skills than does the Elementary. The Intermediate has a much higher vocabulary level than does the Elementary, and it also has much longer paragraphs. It is our view that these factors, the points raised in Section 2, and the grade effects discussed in this section and in Section 3.7 warrant abandoning the MAT standard scores and using the raw scores instead.

In the next section, we discuss the dependent variable used for our raw score analyses.

6.3.2 Definition of the Dependent Variable (Y_{ijkm})

The decision to abandon the standard score metric caused two problems. The first was that the proportion of correct responses is not a dependent measure with constant variance, as the standard scores (in theory) nearly are. Consequently, we could not expect our least squares procedures to work well on this measure. The second was that, where we did not give the same pre- and post-test, we had to rescore the MAT using only parallel items. The items we judged parallel are shown in Appendix E. Table 6-4 shows the number of parallel items we used, by grade, and the total number of items possible. We selected items based on our judgment of what was parallel, not based on the MAT publisher's intentions. Unfortunately, we did not find many parallel items where we did not administer identical pre- and post-tests.

We imposed an additional restriction on the items at grades 4 and 8, where we did administer the same pre- and post-tests. As reported in Appendix G, we had determined a subset of MAT items that we were reasonably sure had been covered at those grades. Consequently, we excluded other items, leaving only those shown in Table G-5 for analysis.

To correct for variability in the variance of students' scores, we weighted each score inversely to its standard deviation. Thus, we let Y_{ijkm} in Eq. 6-1 be:

$$T(P_{ijkm}) = \frac{\sqrt{N_{ijm}}\, P_{ijkm}}{\sqrt{P_{ijkm}(1 - P_{ijkm})}} ,$$

where P_{ijkm} is the proportion of items that student ijm answered correctly at testing k, out of the N_{ijm} possible. When P = 0, we set T(P) to its smallest possible finite value for N. When P = 1, we set T(P) to its largest possible finite value for N.

Table 6-4

NUMBER OF ITEMS ANALYZED AND NUMBER OF POSSIBLE ITEMS, BY GRADE

Grade	Test		Total Reading		Total Math	
	Fall	Spring	Number Parallel	Number Possible	Number Parallel	Number Possible
1	Primer	Primary I	6	72*	24	34
1 (Canton)	Primer	Primer	72	72	34	34
2	Primary I	Primary II	10	77	33	62
3	Primary II	Elementary	13	84	46	108
4	Elementary	Elementary	78	95	58	115
5	Elementary	Intermediate	12	95	41	115
6	Elementary	Intermediate	12	95	41	115
7	Intermediate	Advanced	11	95	46	115
8	Advanced	Advanced	73	95	38	115
8 (R-3)	Advanced	Advanced	73	95	105	115

*The Word Analysis subtest in the MAT Primary I corresponds to the Listening for Sounds subtest in the MAT Primer. The Reading subtest in the Primer corresponds to both the Word Knowledge and Reading subtests in the Primary I, but the correspondence of the latter subtests with the Reading subtest was too difficult to analyze.

The transformation has the effect of changing proportions according to the following tabulation:

P	$T(P)/\sqrt{N}$
0.1	0.33
0.2	0.50
0.3	0.65
0.4	0.82
0.5	1.00
0.6	1.22
0.7	1.53
0.8	2.00
0.9	3.00

The transformed values increase faster for larger proportions than for smaller, making it somewhat easier to detect small differences in large proportions than to detect the same differences in small proportions. Thus, whether the transformation works in favor of noting differences, or against, depends on the size of the proportions involved.

Any artifacts introduced by the transformation are secondary to the artifacts that are introduced by having so few items to analyze after the elimination of those that are not parallel pre and post. We were surprised to discover that the tests had so few parallel items, given that the tests can supposedly be used in or out of level.

In the next sections, we report the results of our analyses of the test results; the items analyzed were selected and transformed as discussed in this section.

6.4 Results of the Analyses of the Unadjusted Transformed Raw Scores

Because of the small number of items analyzed, it is inappropriate to regard the analyses of the transformed raw scores as definitive. These are, however, the best data on student achievement we have.

Table 6-5 shows the unadjusted "effects" of the responsiveness and implementation variables described in Section 6.2.* Tabled are the average gains in students' transformed raw scores for teachers that site visitors scored as having good responsiveness, and the average gains for the students of those scored as having bad responsiveness. Similarly, tabulations are shown for implementation--that is, for students of teachers judged to have well or poorly implemented projects. When we could not reasonably separate the implementation and responsiveness ratings, we combined them as discussed in Section 6.2. These results are labeled as "good/well" or "bad/poor."

Our two main conclusions are that:

- About 80% of the time, students of teachers identified as having good responsiveness generally showed higher average gains than did students of teachers identified as not responsive.
- About 50% of the time, students of teachers identified as having well-implemented projects had larger average gains than students of teachers having poorly implemented projects.

---

*Raw data files were prepared by John Rollin; analyses were executed by George Black and Pat McCall.

Thus, responsiveness as defined in our observations seems a fairly good predictor of which teachers' students will show larger average gains on our transformed metric, while our implementation rating does not do as well.

Our results must be viewed with caution, however, because of the uneven distributions of numbers of students in the various categories of teachers, and, as already noted, because of the generally small number of items being analyzed. A further caution is that our "bad" and "poor" teacher categories also included the "so-so" teachers (see Section 4.7 for definitions of these terms).

Table 6-6 shows the corresponding tabulations for the transformed MAT raw scores for math. Here, neither of our observation variables is very successful. However, at the fourth and eighth grades, where we have the best evidence that our dependent variables are relevant, an association exists between observers' judgments of good responsiveness and gains. For the other grades, we find it difficult to say that responsive teachers are successful in teaching MAT reading items but not successful in teaching MAT math items. This difficulty is especially obvious because so few items and students were available for some responsiveness/nonresponsiveness comparisons.

Even without the problems discussed, we would be reluctant to attribute the evident success of "responsive" teachers just to their rated responsiveness; there are competing explanations that we have not investigated. One such explanation is that a bias exists in the age, race, and sex distribution of students, a bias that works in favor of the responsive teachers and against those having well-implemented projects (although these two factors are not independent).

The bivariate regression model described in Section 6.2, if it fits the observations well, will permit a comparison that eliminates these biases. The interpretation of this model is presented in the next section.

6.5 Results of the Analyses of the Adjusted Transformed Raw Scores

To implement the model discussed in Section 6.2 using the reading dependent variables described in Section 6.3, we ran ten stepwise bivariate regressions: one for each grade, 1 through 8; a separate run for PTR Canton, Mississippi, grade one; and a separate run for R-3, grade 8. For mathematics we ran the same regressions, but excluded projects, where mathematics is not part of the curriculum. At HIT and IRIT, "good responsiveness" and "well implemented" are completely confounded. In Catch-Up, grades 2 and 3, these variables are nearly confounded, and we combined them into a single variable as discussed in Section 6.2. The complete specifications of the regression equations for reading and math are shown in Appendix D.

Because we can calculate the difference $B_{12m} - B_{11m}$ without calculating the parameters B_{ijm}, our procedure was to minimize the variance of the residuals:

$$R_{ijkm} = B_{ijm} + \varepsilon_{ijkm} \quad .$$

Our first concern is for the models' fit to the data.

6.5.1 Goodness of Fit for Fall and Spring Regression Runs

To assess the goodness of fit of our model at each grade, we calculated a residual, R_{ijkm}, for each case:

$$R_{ijkm} = B_{ijm} + \varepsilon_{ijkm} \quad ,$$

where ε_{ijkm} is the error for individual ijm at time k and where B_{ijm} is an unestimated parameter associated with each individual. This parameter is introduced because the R_{ij1m} may not be independent of the R_{ij2m}. For each bivariate regression equation, goodness of fit was assessed by examining the joint distribution of the R_{ij1m} and R_{ij2m} as a function of our implementation rating. As expected, the R_{ij1m} were often highly positively correlated with the R_{ij2m}. According to our model, there should be no large negative correlations and there were none for either reading or math, for either implementation status.

Table 6-7 shows the standard deviations of the residuals for the fall and spring transformed reading raw scores. If the equations fit well at each implementation status, the standard deviation of the residuals is about 1.

Grade 3 E-3 and grade 1 Canton PTR show rather large standard deviations of reading residuals, which indicates a poor fit to those data. Overall, reading data for the other grades are fit reasonably well, but the fit within grade does vary by implementation status, fall or spring regression equation, and PIP. At the second and third grades, Conquest's fit is slightly worse than Catch-Up's, but the reverse is true at the fourth grade. In IRIT the data of teachers with poor implementation tend to fit better, while in Conquest the data of such teachers are fit somewhat worse.

2

Table 6-5

UNADJUSTED "EFFECTS" OF TEACHER RESPONSIVENESS AND IMPLEMENTATION ON STUDENT GAINS ON THE TRANSFORMED RAW SCORES: READING

	Catch-Up				Conquest			
	Responsive		Implemented		Responsive		Implemented	
Grade	Good	Bad	Well	Poorly	Good	Bad	Well	Poorly
Grade 1								
Mean	2.512	2.693	2.119	2.823				
N	8	8	5	11				
Grade 1 (Canton)								
Mean								
N								
Grade 2	Good/Well		Bad/Poor					
Mean	0.812		0.048		3.056	0.985	1.889	1.972
N	6		9		26	31	29	28
Grade 3								
Mean	0.243		0.481		2.029	1.161	1.596	1.244
N	7		8		13	21	24	10
Grade 4								
Mean	1.649	1.069	1.232	1.685	2.252	1.569	2.041	1.861
N	24	14	21	17	28	27	17	38
Grade 5								
Mean	0.534	0.505	0.423	0.614	1.164	1.335	0.683	1.514
N	23	20	21	22	19	14	11	22
Grade 6								
Mean	0.828	-0.035	0.935	0.086	2.397	2.034	1.256	2.598
N	17	6	14	9	18	15	9	24
Grade 7								
Mean								
N								
Grade 8								
Mean								
N								
Grade 8								
Mean								
N								

	HIT				IRIT			
	Responsive		Implemented		Responsive		Implemented	
Grade	Good	Bad	Well	Poorly	Good	Bad	Well	Poorly
Grade 1								
Mean								
N								
Grade 1 (Canton)								
Mean								
N								
Grade 2								
Mean								
N								
Grade 3					Good/Well		Bad/Poor	
Mean					2.509		1.612	
N					13		52	
Grade 4								
Mean					1.871		-0.377	
N					28		6	
Grade 5								
Mean								
N								
Grade 6	Good/Well		Bad/Poor					
Mean	0.581		0.050					
N	28		12					
Grade 7								
Mean	0.977	0.887	0.837	1.158				
N	8	55	51	12				
Grade 8	Good/Well		Bad/Poor					
Mean	0.940		0.718					
N	52		9					
Grade 8								
Mean								
N								

	PTR				R-3			
	Responsive		Implemented		Responsive		Implemented	
Grade	Good	Bad	Well	Poorly	Good	Bad	Well	Poorly
Grade 1								
Mean	0.731	1.567	1.202	1.422				
N	17	37	29	25				
Grade 1 (Canton)								
Mean	6.077	6.853	--	6.631				
N	18	45	--	63				
Grade 2								
Mean								
N								
Grade 3								
Mean								
N								
Grade 4								
Mean								
N								
Grade 5								
Mean								
N								
Grade 6								
Mean								
N								
Grade 7								
Mean								
N								
Grade 8								
Mean								
N								
Grade 8								
Mean					2.014	1.222	--	1.664
N					288	228	--	516

Note: "Effects" comparable to Eq. 6-2 are the differences between well and poor, or between good and bad. Mean = average fall-spring gain in students' transformed raw scores; N = number of students.

Table 6-6

UNADJUSTED "EFFECTS" OF TEACHER RESPONSIVENESS AND IMPLEMENTATION ON STUDENT GAINS ON THE TRANSFORMED RAW SCORES: MATH

	Catch-Up				HIT				R-3			
	Responsive		Implemented		Responsive		Implemented		Responsive		Implemented	
Grade	Good	Bad	Well	Poorly	Good	Bad	Well	Poorly	Good	Bad	Well	Poorly
Grade 1												
Mean	0.105	1.070	1.075	0.393								
N	7	7	4	10								
Grade 2	Good/Well		Bad/Poor									
Mean	4.536		0.654									
N	5		8									
Grade 3												
Mean	-0.124		0.305									
N	7		16									
Grade 4												
Mean	3.401	2.111	3.186	2.487								
N	18	14	16	16								
Grade 5												
Mean	0.708	1.049	0.158	1.565								
N	22	18	20	20								
Grade 6												
Mean	0.505	2.286	0.491	1.772	--	0.490*	--	0.490*				
N	18	7	15	10	--	23	--	23				
Grade 7												
Mean					1.370	0.562	--	1.233				
N					39	8	--	47				
Grade 8												
Mean					--	1.034*	--	1.034*				
N					--	40	--	40				
Grade 8												
Mean									1.246	1.217	4.361	1.225
N									269	291	1	559

Note: "Effects" comparable to Eq. 6-2 are the differences between well and poor, or between good and bad. Mean = average fall-spring gain in students' transformed raw scores; N = number of students.

*Responsiveness and implementation are completely confounded.

Throughout Table 6-7 are instances in which the standard deviation of the residuals for students having teachers with poorly implemented projects is two to three times that of students having teachers with well-implemented projects, and vice versa. However, with the exception of grades 4 and 8, the standard deviations at all grades are less than 3.5 times their expected value, if we exclude Canton PTR.

Based on our analysis of reading residuals, we conclude that our model for reading scores is not adequate for R-3 at grade 8 or for PTR at Canton. Grade 4 shows generally higher residuals than do the other grades, especially in Catch-Up. The equations do not fit students in the various implementation categories equally well, but the differences are not as great as those seen between grades.

Table 6-8 shows the standard deviations of the residuals for the math equations. As with reading, eighth grade R-3 stands out as poorly fit, with seventh grade HIT also showing large standard deviations. At the other grades, our models fit the math data fairly well, except at sixth grade Catch-Up, where children of teachers with poor implementations are not fit as well as are children of teachers who have well-implemented projects.

In summary, we have reasonably good fits to both reading and math data at all grades except grades 1 and 8. At grade 8, as discussed in Section 5, the MAT was not especially relevant to the PIP curriculum, so we will not try to find a better model. The Canton data have large variances, possibly because of the uneven implementation of the program, or because of the small number of items being analyzed.

In the next section, we discuss the implications of our model for the assessment of PIP impact on achievement.

6.3.2 Regression Analysis of the Effect of PIP Implementation on MAT Transformed Raw Scores

Our general model (Eq. 6-1) is such that, if we regress fall-spring gains on the independent variables, the resulting coefficients are the differences between the corresponding fall coefficient and spring coefficient. We used this propriety to calculate

$$\bar{B}_{1m} = B_{12m} - B_{11m} \quad \text{and} \quad \bar{B}_{2m} = B_{22m} - B_{21m}$$

Table 6-7

STANDARD DEVIATION OF RESIDUALS FOR FALL AND SPRING TRANSFORMED READING RAW SCORES, BY PIP, GRADE, AND IMPLEMENTATION STATUS

Grade	Catch-Up				Conquest				HIT			
	Poor		Well		Poor		Well		Poor		Well	
	Fall	Spring	Fall	Spring	Fall	Spring	Fall	Spring	Fall	Spring	Fall	Spring
Grade 1												
SD	0.950	1.676	1.131	1.260								
N	11		5									
Grade 1 (Canton)												
SD												
N												
Grade 2												
SD	0.775	0.869	2.653	0.830	1.616	2.276	2.021	2.833				
N	10		5		28		29					
Grade 3												
SD	0.968	1.556	0.833	1.843	1.085	3.062	[illegible]	[illegible]				
N	10		5		10		[illegible]					
Grade 4												
SD	3.010	4.319	3.020	3.245	2.814	3.3[illegible]	[illegible]	[illegible].833				
N	17		21		38		[illegible]					
Grade 5												
SD	0.991	1.075	1.530	1.208	1.025	1.12[illegible]	[illegible]	[illegible]883				
N	22		21		22							
Grade 6												
SD	3.599	2.289	1.631	2.452	2.235	2.756	[illegible]	0.974	0.879*	[illegible].290*	1.635*	2.151*
N	9		14		24		9		1[illegible]		29	
Grade 7												
SD									1.114*	[illegible].243*	1.[illegible]2*	1.75[illegible]
N									1[illegible]		51	
Grade 8												
SD									0.[illegible]5*	1.340*	2.121*	2.647*
N									9		52	
Gr[illegible] 8												
[illegible]												
N												

Grade	IRIT				PTR				R-3			
	Poor		Well		Poor		Well		Poor		Well	
	Fall	Spring	Fall	Spring	Fall	Spring	Fall	Spring	Fall	Spring	Fall	Spring
Grade 1												
SD					0.905	1.415	0.908	1.632				
N					25		29					
Grade 1 (Canton)												
SD					1.085	5.998	--	--				
N					63		--					
Grade 2												
SD												
N												
Grade 3												
SD	1.226*	2.358*	2.710*	2.95[illegible]*								
N	52		13									
Grade 4												
SD	1.207*	1.235*	3.538*	3.61[illegible]*								
N	6		28									
Grade 5												
SD												
N												
Grade 6												
SD												
N												
Grade 7												
SD												
N												
Grade 8												
SD												
N												
Gr[illegible] 8												
[illegible]									[illegible].548	6.298	--	--
N									516		--	

Not[illegible]: N = numbe[illegible] [illegible]

*In these data, "good re[illegible]nsiveness" [illegible] "well impl[illegible]ented" are completely con[illegible]nded.

Table 6-8

STANDARD DEVIATION OF RESIDUALS FOR FALL AND SPRING TRANSFORMED MATH RAW SCORES, BY PIP, GRADE, AND IMPLEMENTATION STATUS

Grade	Catch-Up				HIT				R-3			
	Poor		Well		Poor		Well		Poor		Well	
	Fall	Spring	Fall	Spring	Fall	Spring	Fall	Spring	Fall	Spring	Fall	Spring
Grade 1												
SD	0.984	1.077	0.762	1.886								
N	10		4									
Grade 2												
SD	1.177	1.292	0.217	7.524								
N	9		4									
Grade 3												
SD	1.171	1.663	2.817	2.540								
N	18		5									
Grade 4												
SD	2.143	4.234	1.705	3.569								
N	16		16									
Grade 5												
SD	1.433	2.065	2.364	1.968								
N	20		20									
Grade 6												
SD	3.429	4.866	2.551	2.065								
N	10		15									
Grade 7												
SD					1.753	2.487	--	--				
N					47		--					
Grade 8												
SD					3.031	3.922	--	--				
N					40		--					
Grade 8												
SD									4.840	5.925	0.0	0.0
N									559		1	

Note: N = number of students.

for reading and math. $\overline{B}_{1m}$ is the "adjusted effect" of a teacher with a well-implemented project, and $\overline{B}_{2m}$ is the adjusted effect of a responsive teacher. For those cases in which we could not estimate $\overline{B}_{1m}$ and $\overline{B}_{2m}$ separately, we estimated a single coefficient $\overline{B}_m$, as discussed in Section 6.2; this coefficient is labeled "well/good" in our tables. Computations were done using the stepwise regression procedure of the "Statistical Package for the Social Sciences" (Nie et al., 1970), so that we would get some idea of the relative importance of $\overline{B}_m$, or $\overline{B}_{1m}$ and $\overline{B}_{2m}$, given the other variables in the equations. Ultimately, we forced the program to take all independent variables shown in Eq. 6-2.

Table 6-9 summarizes the process for gains on our transformed reading score, and Table 6-10 summarizes the run for gains on our transformed math scores.

In Table 6-9, the final coefficient of determination is not particularly impressive at any grade, especially at Canton PTR and grade 7 HIT. To some extent, these results could be anticipated from the results of our analyses of residuals. However, the coefficient of determination for grade 7 HIT is worse than anticipated.

In this table, we have shown the sign of the final value of the coefficient of $\overline{B}_m$, or of $\overline{B}_{1m}$ and $\overline{B}_{2m}$, as a coefficient to the increase in the coefficient of determination due to $\overline{B}_m$, or to $\overline{B}_{1m}$ and $\overline{B}_{2m}$, at the time it entered the equation. In parentheses, we show the order in which $\overline{B}_m$, or $\overline{B}_{1m}$ and $\overline{B}_{2m}$, entered the stepwise procedure. Generally, the implementation and responsiveness parameters do not add much to R^2; that is, the effect of these variables is small.

Comparing the signs of the effects shown in Table 6-9 with the corresponding signs calculated from Table 6-5 shows that adjusting for age, race, and sex alters the picture of responsiveness and implementation impacts. Seven effects that were positive in the table of unadjusted effects are negative effects after adjustment. Four effects that were negative before adjustment are positive afterwards.

However, examination of the increase in the coefficient of determination due to inclusion of $\overline{B}_m$, or $\overline{B}_{1m}$ and $\overline{B}_{2m}$, shows that in five PIP grade combinations the increase is less than 0.0015. In these cases, knowledge of implementation status or teacher responsiveness does not add information not present in the variables in the equation before the entry of $\overline{B}_m$, or $\overline{B}_{1m}$ and $\overline{B}_{2m}$.

If we ignore cases in which $\overline{B}_m$, or $\overline{B}_{1m}$ and $\overline{B}_{2m}$, add this little to the coefficient of determination, in three cases positive unadjusted

Table 6-9

INCREASE IN THE COEFFICIENT OF DETERMINATION FOR SPRING-FALL DIFFERENCES IN READING

Grade	Number: Class	Number: Steps	[illegible]: Responsive	[illegible]: Well Implemented	Compost: Responsive	Compost: Well Implemented	[illegible]: Responsive	[illegible]: Well Implemented	[illegible]: Responsive	[illegible]: Well Implemented	EIR: Responsive	EIR: Well Implemented	R-3: Responsive	R-3: Well Implemented	Final R^2
1	20	7	+0.012 (1)	[illegible] ([illegible])							+0.051 (2)	+0.010 (6)			0.214
1 (Canton)	61	3									+0.006 (1)	--			0.005
2	72	5	Well/Good +0.0[illegible] ([illegible])		+0.191 ([illegible])	[illegible] ([illegible])									0.218
3	114	[illegible]	-0.015 (2)		+0.00[illegible] (3)	-0.000 (7)			Well/Good +0.022 (1)						0.052
4	127	[illegible]	+0.038 (4)	+0.023 ([illegible])	+0.025 (1)	+0.001 (7)			+0.009 (3)						0.095
5	76	2	-0.008 (1)	-0.011 (1)	+0.010 (5)	-0.008 (4)									0.032
6	1[illegible]	9	+0.060 (5)	+0.001 (6)	+0.059 ([illegible])	-0.015 (2)	Well/Good +0.000 (7)								0.117
7	53	5					-0.000 (3)	+0.000 (4)							0.012
8	61	4					Well/Good -0.010 (3)								0.102
8	516	5											+0.004 (3)	--	0.057

Note: Numbers in parentheses indicate the order entered; sign is the sign of the coefficient in the final regression equation.

Table 6-10

INCREASE IN THE COEFFICIENT OF DETERMINATION FOR SPRING-FALL DIFFERENCES IN MATH

Grade	Number Cases	Number Steps	Catch-Up Responsive	Catch-Up Well Implemented	HIT Responsive	HIT Well Implemented	R-3 Responsive	R-3 Well Implemented	Final R^2
1	14	2	-0.093 (1)	+0.038 (2)					0.131
2	13	1	Well/Good +0.197 (1)						0.197
3	23	1	-0.015 (1)						0.015
4	32	2	+0.058 (1)	-0.001 (2)					0.059
5	40	2	+0.001 (2)	-0.188 (1)					0.189
6	25	2	-0.111 (1)	-0.020 (2)					0.131
7	47	1			+0.027 (2)	--			0.027
8					Not run				--
8	560	5					+0.005 (2)	--	0.027

Note: Numbers in parentheses indicate the order entered; sign is the sign of the coefficient in the final regression equation.

effects are negative after adjustment and in three cases negative unadjusted effects become positive.

If we double-count $\overline{B}_{TR}$ as both well implemented and responsive, in four cases the adjusted effect of responsiveness is negative and in ten cases it is positive. In seven cases the adjusted effect of being judged well implemented is negative and in four it is positive.

Therefore, even after adjusting for age, race, and sex, our general conclusions concerning the unadjusted effects hold. If implementation or responsiveness has any impact, good responsiveness, as we have defined it, is associated with small gains on student achievement tests in a variety of PIP and grade combinations.

Table 6-10 shows corresponding statistics for math. Again, the equations for grades 7 and 8 stand out as not being very good predictors of the values of fall-spring gains. In the math data the effect of adjustment was to convert one unadjusted effect from positive to negative, and one from negative to positive. Thus, adjustment for age, race, and sex does not much alter our conclusions concerning the effects of PIP implementation and teacher responsiveness on our math variable. However, the adjustment did increase the number of positive effects for responsiveness from four to five and the number of negative implementation effects from two to three.

Therefore, in our formal analyses, teacher responsiveness--more often than implementation of PIP philosophy and specifications--is associated with increases in the transformed raw scores.

Caution should be exercised in generalizing our results to the unobserved children in our study or to future studies that use procedures different from ours. Nevertheless, our result on achievement is fairly clear: As judged by the norm-referenced analyses and by the regression models just reported, implementation of the PIP philosophy and procedures did not raise MAT scores to any impressive degree. However, the MAT content was not especially relevant to PIP curricula, and the MAT standard scores may not be well-suited to valid norm-referenced procedures.

7 SYNOPSIS AND RECOMMENDATIONS

7.1 Summary

In this section we review the outcomes of our evaluation summarizing our conclusions.

The main issue addressed was the validity of the PIP replication principle and the associated norm-referenced analysis. Based on our evaluation principles, we conclude that the replication principle was false: there is little reason to believe that packages of the type we evaluated would make MAT scores dramatically increase.

To reach this conclusion we examined several peripheral issues concerning norm-referenced analyses and the one-third standard deviation criterion of educational significance. We pointed out several technical flaws in the norm-referenced "t" test, and showed that, as applied to the MAT, the criterion of educational significance was not a constant proportion of expected growth. In this sense the criterion was not equally stringent at all grades.

Examining our data relative to the MAT norms, we noticed that in the fourth grade it was more difficult to reach criterion than in the third and fifth grades. We executed computer simulations which confirmed the trend. We conclude that there is some artifact in the published norms at this grade.

Based on our curriculum analyses and site visits we found that the PIPs did induce projects which were adequate copies of what was packaged; however what was packaged was not sufficient to implement the same curricula across sites.

Our analysis of the correspondence between the MAT and the curriculum materials which were both listed in the PIP and used in the projects provided evidence that, except at the lower grades, MAT items were not sensitve to such materials.

Finally, in our least squares analysis, we found that teacher responsiveness was more often associated with gains in test scores than was good implementation of PIP specifications.

7.2 Methodological Recommendations

The preceeding sections have shown the utility of approaching evaluation through the use of a principle of description: the evaluator must display a connection between the outcomes of interest and the treatment. This has implications for:

1) The way USOE (or others) should decide that a project is effective, successful or exemplary. That is, the connection between the data offered as evidence of effectiveness and the content and procedures of the project should be judged for its reasonableness.

2) The way in which PIP-type packages are created in the future. A project could be analyzed from its outcomes, backwards to the proximal events which could have caused them, (all the way back to the management strategies which promoted such events if desired). The information in the package might then be more likely to convey the effective elements.

3) The way evaluations are conducted. That is, to evaluate the effects of a treatment appropriately would require a description of the treatment at the level of discourse relevant to the effects examined.

Our application of this idea to the evaluation of Project Information Packages led us to examine one standardized test and the associated norm-referenced analysis in detail. It was concluded that there were probably defects in that tests's fourth grade norms. We also found through our simulations that "equipercentile growth" could be achieved by guessing alone. We developed evidence that, in this study at least, compensatory education teachers do not teach to the MAT, except perhaps at the first and second grade levels. The test items however were sensitive to the responsiveness of the teacher, where responsiveness was judged by trained observers and directly coded for regression analyses. Similar analyses might be fruitful in other evaluations which use achievement test scales.

Based on our results, we would recommend not using the MAT standard scores as the principle measure of project success. We would also recommend that the consumer of standardized tests not be drawn into the belief that standardized achievement tests are equivalent, even if tests like the Anchor Test Study claim to display "equivalence" for some tests.

We recommend that the implications of assuming that any test's cross-sectional norms are longitudinally valid be seriously considered before the norms are used out of level. We also question whether there is really a single trait called "Reading Achievement," and we question whether we know how it grows from the first grade to the twelfth. If there is no such trait or we do not know its laws of growth, then our achievement scale is nugatory.

It seems to us that the trend of these considerations is to abandon norm-referenced, standarized tests with their simple scales. What is needed are tests with items that are sensitive to those skills we are taught. Then if it were determined that teachers were actually teaching such skills such tests would form a convenient foundation for uniformly evaluating diverse projects.

Appendix A

MANUAL OF PROCEDURES FOR PROJECT INFORMATION PACKAGES TESTING

Table of Contents

April 1976

Spring 1976

Manual of Procedures for

Project Information Packages Testing

Project Information Packages
Evaluation Study

SRI Project URU-3556

INTRODUCTION

The purpose of this manual is to explain the tasks and responsibilities of field data collectors who will be conducting the Spring 1976 testing program as part of the evaluation of Project Information Packages (PIPs). Recognizing the variations between PIPs and projects, an effort has been made to provide detailed directions to assure uniform testing procedures across all PIP projects.

Uniformity is important for two reasons. First, for purposes of standardization, it is important that testing procedures and conditions approximate, as closely as possible, those described by the test authors and publishers. Second, in order to provide reliable results, it is important to administer the tests as consistently as possible to different groups of students both within and across projects. Thus, it is important that field data collectors involved in the testing program thoroughly understand their tasks and responsibilities prior to assuming them and that they adhere to the guidelines for performing those tasks throughout the testing period.

The field data collection staff will consist of the SRI Test Supervisor, the local Site Assistant and, where necessary, additional local personnel to serve as Testers and Monitors. The SRI Test Supervisor will assume overall responsibility for the testing program. Where possible, the SRI Test Supervisor will assume the role of Tester with the local Site Assistant serving as Monitor. In projects with large numbers of students to be tested, the SRI Test Supervisor will hire, train, and supervise local personnel as Testers and Monitors. Car will be taken to select people who do not have a vested interest in the project. Local personnel who cannot demonstrate performance skills required by SRI, during on-site training, will not be utilized as a Tester or Monitor.

The local Site Assistant will be responsible for completing all necessary preparations for testing, will assist during testing, and will be responsible for returning all test materials to SRI following completion of testing.

This manual is divided into four sections.

Section I - Tests and Test Sample

Section II - Site Assistant's Guide to Testing Preparations

Section III - Tester's and Monitor's Responsibilities

Section IV - Site Assistant's Instructions for Returning Test Materials to SRI

Section I

TESTS AND TEST SAMPLE

The test battery for each student will consist of three types of tests:

1) The Metropolitan Achievement Test (MAT)

2) One of two affective tests

   a) For first and second graders the FACES Attitude Inventory, or
   b) For third through ninth graders the Intellectual Achievement Responsibility Scale (IAR)

3) A PIP and site-specific student attitude questionnaire

   a) For first and second graders, one which uses the FACES format, or
   b) For third through ninth graders, one which uses the Coopersmith format.

Metropolitan Achievement Test

The Metropolitan Achievement Tests (MATs) are a series of measures designed to tell how much pupils have learned in important content and skill areas of the school curriculum.

There are six levels of the MAT. The levels that will be used in April for each grade in the PIP evaluation are as follows:

Grade 1	Primary I
Grade 2	Primary II
Grade 3	Elementary
Grade 4	Elementary
Grade 5	Intermediate
Grade 6	Intermediate
Grade 7	Advanced
Grade 8	Advanced
Grade 9	Advanced

There are three forms of the test (F, G, and H) at each level. Only form F will be used. Each member of the Test Battery has several subtests.

FACES Attitude Inventory

The FACES Attitude Inventory is designed to gather information about the student's general feeling toward himself, toward others, toward school, and learning in general. In response to themes pictorially presented in the test items and verbally described by the Tester, each student shows his feelings by marking one of three responses: a happy face, a so-so face (not happy, not sad), or a sad face. There are 14 items in the FACES Attitude Inventory.

Intellectual Achievement Responsibility Scale (IAR)

The Intellectual Achievement Responsibility Scale (IAR) is aimed at assessing a student's belief in reinforcement responsibility in academic achievement situations.

The IAR scale consists of 20 forced-choice items. Oral presentation will be made by the Tester to students in grades three through five. Students in grades six through nine will be administered the IAR in written form.

Student Attitude Questionnaires

The Student Attitude Questionnaires are designed to assess student feelings toward the PIP projects. The Student Attitude Questionnaires are PIP-specific and, in some instances, site-specific. The FACES format will be given to students in grades one and two, and the Coopersmith Self-Esteem Inventory will be given to students in grades three and up.

Oral presentation will be made by the Tester to students at all grade levels. The Student Attitude Questionnaires should be administered immediate following the affective tests.

Test Sample

The test sample in April will include only those students tested in the Fall for whom we have valid test data. The IRIT sample will consist of middle cycle students tested in the Fall.

No additional students will be tested even though they are in the program.

Table 1 shows the grade levels that will be tested at each project as well as the subject areas and affective tests each grade level will receive.

Table 1

PIP TEST PLAN

PIP	Project	METROPOLITAN ACHIEVEMENT TESTS: Primary I Grade 1 Read	Primary I Grade 1 Math	Primary II Grade 2 Read	Primary II Grade 2 Math	Elementary Grade 3 Read	Elementary Grade 3 Math	Elementary Grade 4 Read	Elementary Grade 4 Math	Intermediate Grade 5 Read	Intermediate Grade 5 Math	Intermediate Grade 6 Read	Intermediate Grade 6 Math	Advanced Grade 7 Read	Advanced Grade 7 Math	Advanced Grade 8 Read	Advanced Grade 8 Math	Advanced Grade 9 Read	Advanced Grade 9 Math	AFFECTIVE TESTS: FACES Grades 1-2	I A R Grades 3-5	I A R Grades 6-9	STUDENT ATTITUDE QUESTIONNAIRES: FACES Grades 1-2	Coopersmith Grades 3-9
Catch-Up	Bloomington, Ind.	x	x	x	x	x	x	x	x	x	x									x	x		x	x
	Brookport, Ill.	x	x	x	x	x	x	x	x	x	x	x	x							x	x	x	x	x
	Galax, Va.	x	x	x	x	x	x	x	x	x	x	x	x							x	x	x	x	x
	Providence Forge, Va							x	x	x	x	x	x								x	x		x
	Wayne City, Ill.	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x			x	x	x	x	x
Conquest	Benton Harbor, Mich.			x		x		x		x		x								x	x	x	x	x
	Cleveland, Ohio	x		x		x		x		x		x								x	x	x	x	x
	Gloversville, NY			x		x		x		x		x								x	x	x	x	x
HIT	Lexington, Miss.											x	x	x	x	x	x	x	x			x		x
	Olean, NY													x	x	x	x	x	x			x		x
CHIT	Bloomington, Ind.					x		x													x			x
	Oklahoma City, Ok.					x															x			x
	Schenectady, NY					x		x													x			x
PTR	Canton, Miss.	x																		x			x	
	Dallas, Texas	x																		x			x	
R-3	Charlotte, N.C.															x	x					x		x
	Lake Village, Ark.															x	x					x		x
	Lorain, Ohio															x	x					x		x
	Schenectady, NY															x	x					x		x

Section II

SITE ASSISTANT'S GUIDE TO TESTING PREPARATION

An Overview

As Site Assistant, you will be responsible for all testing preparations. You should:

- Check Test Rosters for accuracy and fill in missing information.
- Prepare Test List worksheets when necessary.
- Prepare a Tester Log for each Test Roster or Test List worksheet.
- Get the testing schedule approved.
- Identify the location of students to be tested and enter teacher names on Test Rosters or Test List worksheets.
- Locate testing areas and furniture.
- Prepare school map. Duplicate for Testers and Monitors.
- Duplicate approved test schedule. Distribute to teachers concerned and principal. Reserve copies for Testers and Monitors.
- Locate space for local training session, if necessary.
- Receive test booklets and check supply against local needs.
- Label test booklets and group booklets with appropriate Test Rosters (or Test List worksheets) and the Tester Logs.

Test Roster--What is it

The Test Roster is an alphabetical listing of students, by school and grade, who were tested in Fall 1975. The Test Roster is furnished by SRI. The Test Roster will serve as the primary source of information in grouping children for testing, labeling booklets, and keeping a record of tests administered.

The format of the Test Roster is described below. A sample Test Roster is provided in Exhibit I.

PROJECT:	Project number (assigned by SRI), name of the PIP and its location (city).
SCHOOL:	School number (assigned by SRI), and name of the school in which students were located in Fall 1975.

TEST GROUP:	Test group identifier--a letter such as A, B, C, etc. or Z (assigned by SRI).
GRADE:	Grade level of students at the time of Fall 1975 testing.
TEACHER NAME:	Blank column to be filled in according to instructions provided in this manual.
PUPIL NAME:	Names of students tested in Fall 1975.
BIRTHDATE:	Birthdates of students listed.
ETH:	Ethnicity of students listed.
SEX:	Sex of students listed.
TUTOR/TUTEE CODE:	Identifies student as a tutor (1) or tutee (2). This column and the information in it is applicable <u>only</u> to the HIT sites.
VALID FALL 75 TESTS:	Identifies those students who have valid test data (YES) from Fall testing and those who have no valid test data (NO) from Fall testing. <u>ONLY STUDENTS WITH VALID FALL TEST DATA ARE TO BE TESTED IN APRIL</u>.
ABSENT 20% OR MORE:	Five digit number for each student listed. (First two digits represent the project number and will be the same for all students in the project. The last three numbers represent unique student ID numbers assigned in the Fall). This column is to be completed according to directions in this manual.
TEST SITTING:	These three columns will be blank and are to be completed by the Tester.
SEE TESTER LOG:	To be completed by the Tester.
COMMENTS:	To be completed by the Tester.

Test Roster--How to use it

Correcting and Completing the Test Roster

When you receive the Test Roster, the first thing you should do is verify the accuracy of the student names, birthdates, and tutor/tutee codes (if applicable). Make corrections on the Test Rosters by crossing out the incorrect information and entering the correct information above it. All blank spaces should be filled in with the appropriate information.

Exhibit 1

Sample Test Roster

DATE RUN: 03/16/76 PAGE: 22

SRI PIP EVALUATION
SPRING 1976 TEST ROSTER

PROJECT: 18 HIT SAN FRANCISCO

SCHOOL: 1 SUNNYVALE
TEST GROUP: Z
GRADE: 6

TEACHER NAME	PUPIL NAME LAST NAME FIRST	BIRTH DATE	ETH	SEX	TUTOR-TUTEE CODE	VALID FALL75 TESTS	CIRCLE IF ABSENT 20 % OR MORE	TEST SITTING .1 .2 .3	SEE TESTER LOG.	COMMENTS
	1 APPLE JOHN	06/19/63	B	M	1	YES	32500			
	2 BEAR MARK	06/05/63	B	M	2	NO	32512			
	3 DANIELS WILLIAM	11/07/62	B	M	2	YES	32513			
	4 HARRIS MARY	0?/01/64	B	F	2	YES	32514			
	5 HOWARD ROBERT	08/31/64	B	M	1	YES	32501			
	6 JOHNSON PAUL	12/03/63	B	M	1	YES	32502			
	7 KENNEDY ROSE	06/17/61	B	F	2	YES	32515			
	8 POWELL DAVID	06/05/61	B	M	2	YES	32516			

A-13

You should then contact the project teachers and ask them to identify those students who have been absent 20% or more during their PIP instructional period. Do NOT ask the school secretary or regular classroom teacher for this information. A 20% absentee rate is approximately one day a week or a total of 35 days during the year (10 days absence during the cycle for IRIT project students). You should circle the five-digit number on the Test Roster corresponding to the student's name for any student who has been absent 20% or more in the column titled such.

Filling in the Teacher Names will be explained in the Test Schedule section (page 14) of this manual.

Identifying Test Groups

The grouping of students into test groups has already been set up according to guidelines established by SRI. The test groups are the same as the test groups that were established in the Fall. The guidelines used were:

First graders - testing in groups of 10 or less
Second graders - testing in groups of 15 or less
All other grades - testing in groups of normal class size.

If the TEST GROUP identifier on the Test Roster is:

A through Y - The list represents a test group. It should not be changed. There may be more than one grade in one test group. If so this will be indicated on the Test Roster by your SRI Test Supervisor.

Z - Pupils are grouped by grade within school.

If the TEST GROUP identifier on the Test Roster is a "Z", then the list of student may be used as is, may be combined with other test groups to form a larger test group, or may be divided into smaller test groups. Your SRI Test Supervisor will indicate how the list is to be handled. If any list has to be divided into smaller test groups, see the section Making up Test List Worksheets from Test Rosters below. Where two small groups are combined to make one test group simply staple the two Test Rosters together and they represent the test group.

Test groups are not to be changed without consulting the SRI Test Supervisor.

Making up Test List Worksheets from Test Rosters

As stated previously, the Test Rosters will serve, in most cases, as the only lists of test groups you will need. However, the SRI Test Supervisor may indicate that Test List worksheets must be prepared and will attach an explanation to the Test Rosters explaining how students listed on the Rosters are to be grouped. It is your job then to prepare the Test List worksheets (see Exhibit 2) according to instructions.

You will find that Test List worksheets will be made up only if one or more of the Test Rosters do not represent a test group. In making up the Test List worksheets students from the same class should be grouped together. The Test Roster will serve as the source document and should be kept and returned to SRI with the test bookets after testing has been completed. (Neither the Test Rosters or Test List worksheets are to be discarded. Both will be returned.)

The top portion of the Test List worksheet contains space for recording the project number, name, and city, school number, school name, and grade. All of this information is on the top of the Test Roster and should be transferred to the Test List worksheet exactly as it appears. You will note that you are not to transfer any test group identifier. The name of the Test Administrator and the date of testing will be filled in by the Tester on the day of testing.

Exhibit 2

TEST LIST WORKSHEET

PROJECT INFORMATI
PACKAGES EVALUATI
Spring 1976

OJECT NO._______ PROJECT NAME ______________________________ CITY ________________________

SCHOOL NO. _______ SCHOOL NAME ____ GRADE _____

TEST ADMINISTRATOR ____ DATE _________________

Teacher Name	I.D. No.	NAME Last	First	ETHNICITY	SEX	SITTINGS 1	2	3	See Test Log
		1.							
		2.							
		3.							
		4.							
		5.							
		6.							
		7.							
		8.							
		9.							
		10.							
		11.							
		12.							
		13.							
		14.							

Ethnicity Codes: B = Black
C = Caucasian
S = Spanish Surname
O = All other

Sex Codes: M = Male
F = Female

The remaining portion of the Test List worksheet contains space for the teacher name, student ID number, name, ethnicity and sex. Filling in the teacher name, will be explained in the Test Schedule section. The other information should be transferred to the Test List worksheet from the Test Roster exactly as it appears.

Please note that only students with valid Fall 1975 tests will be tested in April. Any student with a "NO" appearing in that column on the Test Roster should not have his name transferred to the Test List worksheet. When the Test List worksheets have been completed, please double check to see that every student with a "YES" in the VALID FALL 75 TESTS column has been entered on a Test List worksheet.

The columns headed SITTINGS and SEE TESTER LOG will be completed by the Tester during testing.

Ethnicity and sex will also be verified by the Tester during testing.

Preparing the Tester Log

A Tester Log must be prepared for each test group (see Exhibit 4, page 32). The purpose of the Tester Log is explained in Section III.

The top portion of the Tester Log provides for project number, project name, city, school number, school name, and grade. This information should be transferred from the Test Roster exactly as it appears.

If you must prepare separate Test List worksheets from the Test Roster, then you will fill in the top portion of each Tester Log exactly as you did the Test List worksheet.

When you are through, you will have sets of either Test Rosters and Tester Logs or Test List worksheets and Tester Logs. Each set will be kept with its associated set of labeled test booklets.

The Test Schedule

You will be sent a test schedule prepared by the SRI Test Supervisor.

Getting Approval of the Test Schedule

After you have received the schedule, present it for review and approval to the principal and regular classroom teachers. Please familiarize yourself beforehand with the Test Sittings--Time Requirements description below so that you may answer questions regarding the need for the time requirements scheduled and explain the need to start each test session as scheduled. At some sites the operation must function like clockwork to work in all tests and all groups, giving each the full allotted time.

After the schedule is finalized, duplicate and present copies to all concerned. Reserve enough copies for distribution to the Testers and Monitors.

Identifying Location of Students

While finalizing the test schedule, it is important to find out where the students will be during the scheduled testing times so they can either be picked up for testing or so the Tester and Monitor will know where to go to test them.

Filling in the Teacher Names on the Test Roster or Test List Worksheet

On the Test Roster (or Test List worksheet) enter the name of the teacher who is supervising the students at the time testing is scheduled. This information is necessary so that Testers will know where to go to pick up the students during testing. If all students in the test group have the same teacher, the teacher's name need only be entered once on the Test Roster or Test List worksheet.

Preparing a Diagram of the School

Please prepare a diagram of the schools in which testing occurs. The diagrams should show the testing area(s), the teachers' rooms where students to be tested can be found, bathrooms, and the principal's office. Duplicate enough copies for distribution to Testers and Monitors. The diagrams will facilitate keeping to the prepared test schedules.

Test Sittings--time requirements

Table 2, Test Sittings, is a listing of typical sittings by grade, tests administered and the actual time allotted to each of the tests. The table does not provide for time between tests, time for distributing booklets, time to allow students to settle down before testing begins or for collecting the booklets when testing is over. The schedules you will receive from the SRI Test Supervisor will include approximately ten minutes between test administrations. This means that students should be ready to come to the testing location ten minutes before the tests are actually to begin.

Locate Testing Area and Furniture

As the testing plan is being discussed, inquiries should be made regarding testing space and furniture since this may directly affect the testing schedule itself. The Site Assistant should ask to see the testing area, determine the adequacy of furniture, and make certain that school personnel who generally use the room are informed of its use as a testing area. If the testing area is ordinarily occupied by a school staff member, group, etc., a test schedule should be provided to them.

Space for Local Training

A one or two day training session will be conducted by the SRI Test Supervisor in projects where local personnel are hired to serve as Testers

Table 2

TEST SITTINGS

Grade 1

Primary I - Reading Subtest

Sitting 1 -	What to Do (Practice)	10
	Reading	30
Sitting 2 -	Word Knowledge	15
	Faces	15
	Questionnaire	5

Primary I - Reading and Math

Sitting 1 -	What to Do	10
	~~Reading~~	~~30~~
Sitting 2 -	Word Knowledge word analysis	15 15
	Faces	15
	Questionnaire	5
*Sitting 3 -	Math Concepts	15
	Math Computation	15

Grade 2

Primary II - Reading Subtest

Sitting 1 -	What to Do	10
	Reading	30
Sitting 2 -	Word Knowledge	18
	[redacted] Faces	15
	Questionnaire	5

Primary II - Reading and Math

Sitting 1 -	What to Do	10
	Word Knowledge	18
	Reading	30
Sitting 2 -	Math Computation	18
	Math Concepts	20
*Sitting 3 -	Problem Solving	25
	[redacted] Faces	15
	Questionnaire	5

Grades 3-4

Elementary - Reading Subtest

Sitting 1 -	What to Do	10
	Reading	25
Sitting 2 -	Word Knowledge	15
	IAR	15
	Questionnaire	5

Elementary - Math Subtest

Sitting 1 -	What to Do	5
	Math Computation	35
	Math Concepts	25
Sitting 2 -	Problem Solving	30
	IAR	15
	Questionnaire	5

Elementary - Reading and Math

Sitting 1 -	What to Do	10
	Word Knowledge	15
	Reading	25
Sitting 2 -	Math Computation	35
	Math Concepts	30
*Sitting 3 -	Problem Solving	30
	IAR	15
	Questionnaire	5

* = day 2

Table 2 (continued)

Grades 5-6

ntermediate - Reading Subtest

itting 1 -	What to Do	5
	Word Knowledge	15
	Reading	25
	IAR	10-15
	Questionnaire	5

Intermediate - Math Subtest

Sitting 1 -	What to Do	5
	Math Computation	35
	Math Concepts	25
Sitting 2 -	Problem Solving	25
	IAR	10-15
	Questionnaire	5

Intermediate - Reading and Math

Sitting 1 -	What to Do	5
	Word Knowledge	15
	Reading	25
Sitting 2 -	Math Computation	35
	Math Concepts	25
*Sitting 3 -	Problem Solving	25
	IAR	10-15
	Questionnaire	5

Grades 7-9

dvanced - Reading Subtest

itting 1 -	What to Do	5
	Word Knowledge	15
	Reading	25
	IAR	10
	Questionnaire	5

Advanced - Math Subtest

Sitting 1 -	What to Do	5
	Math Computation	35
	Math Concepts	25
Sitting 2 -	Problem Solving	25
	IAR	10
	Questionnaire	5

Advanced - Reading and Math

Sitting 1 -	What to Do	5
	Word Knowledge	15
	Reading	25
Sitting 2 -	Math Computation	35
	Math Concepts	25
*Sitting 3 -	Problem Solving	25
	IAR	10
	Questionnaire	5

= day 2

and Monitors. The training session will generally be conducted on the Thursday and/or Friday preceeding the week set aside for testing. The number of days set aside for training will be dependent on the number of grade levels being tested. The SRI Test Supervisor will inform the Site Assistant if a training session is to take place and will specify the date(s).

The Site Assistant should arrange for a training location that will comfortably accomodate the participants. The Project Director should be consulted on a training location. After the training location has been arranged for, the Site Assistant should arrange to have all testing materials moved to the training location on the day training takes place and see that furniture is adequate.

<u>Receive Test Booklets</u>

Test materials required for each project are shipped from SRI to either the Site Assistant or to the Project Director. If the test materials are not sent directly to the Site Assistant, the SRI Test Supervisor will notify the Site Assistant as to where the materials were sent.

Upon receipt of the test materials, the Site Assistant should open the test cartons and check the contents against the Packing Invoice (see Exhibit 3) which will appear in one of the test cartons. Care should be exercised in opening the cartons so that they can be used for the return shipment.

Upon examining the Packing Invoice the Site Assistant will notice that the top portion of the invoice, as well as columns 1 and 2, will have been completed at SRI. The Site Assistant should count the number of tests received for each level that is recorded in column 1 and enter the number received in column 3. Columns 2 and 3 should agree. The Packing Invoice

PROJECT INFORMATION
PACKAGES EVALUATION
Spring 1976

Exhibit 3

PACKING INVOICE

PROJECT ______________________ PROJECT NUMBER __________

Entries are made on the Packing Invoice at SRI and at the local project.

-- Columns 1 & 2 are completed at SRI before the tests are shipped to the local project.

-- Columns 3 through 6 are completed at the local project by the Site Assistant.

-- Column 7 is completed at SRI after the tests have been received back at SRI.

This Packing Invoice must be returned to SRI with the test booklets.

--Columns--

1	2	3	4	5	6	7	8
Tests Name / Level	Number Shipped From SRI	Number Recv'd on Site	Number Used	Number Not Used	Number Return-ed	Number Recv'd at SRI	Comments
Tester Kits							

should then be checked against Table 1, page 6 of this manual, to insure that the correct tests have been received. Any inconsistencies should be brought to the immediate attention of the SRI Test Supervisor. If the Test Supervisor cannot be reached, a collect call should be made to Nancy Craig, (415) 326-6200, ext. 2995.

After testing has been completed, columns 4, 5, and 6 of the Packing Invoice should be completed by the Site Assistant. Column 4 "Number Used" refers to the number of test booklets used by students, and column 5 "Number Not Used" refers to the number of blank booklets (this includes booklets which have labels on them but were not administered). Column 7 is reserved for SRI use and will be completed at SRI as the returned tests are inventoried and logged in.

Labeling Test Booklets

After the test carton(s) have been inventoried, test booklets should be labeled and grouped for each day's testing. Each student's test booklets must be labeled before they are presented to him.

The following procedures should be followed in the labeling process:

1. Determine which group of students will be tested first and select the corresponding Test Roster (or List).

2. Count the number of students listed on the Test Roster (or List). Count out an equal number of the appropriate MAT and affective test booklets (IAR and FACES), and Student Attitude Questionnaires (FACES and Coopersmith).

3. For each name on the Test Roster (or Test List worksheet), select the corresponding printed peel-off student label generated at SRI. Place the label in the top right-hand corner of the front cover of each test booklet. This label is to remain on the test. The student's name is not to be recorded in any other space on the booklet.

4. Review each Test Roster for corrections that were made and make appropriate corrections on the student labels.

Step 3 is to be repeated for each student on the Test Roster (or Test List worksheet). It is suggested that the MAT, affective test and attitude questionnaire be labeled at the same time for each student to reduce the possibility of error. The test booklets should be grouped and banded with the corresponding Test Roster (or Test List worksheet) in preparation for testing. At sites where there will be several Testers, the Site Assistant should group tests by school, test team, and test group or hour of test administration.

Section III

TESTER'S AND MONITOR'S RESPONSIBILITIES

The Night Before Each Day's Testing

The Tester must check the sets of test booklets to be used the next day against the individual Test Rosters (or Lists) to make sure that the right number of booklets are available for each grade level and that the names on the booklets match the names on the Rosters (or Lists). The sets of booklets should be put in order of use according to the next day's schedule. The list of required materials should be checked and materials assembled. See the list of materials below. Note that scratch paper, for instance, is required for some math subtests but not allowed for others.

At the Beginning of Each Day's Testing

The Tester and Monitor should arrive at the school early, check in with the principal or school secretary as prescribed by school policy, and proceed to the testing location. They should have with them the following materials:

- Testing schedule
- Test booklets to be used that day
- Test Rosters (or Lists) and Tester Logs for each set of booklets
- Appropriate test administration booklets
- Map of school room locations
- Sharpened pencils for each student plus extras in case of broken leads
- Scratch paper for math subtests where allowed
- "Testing - Do Not Disturb" sign
- Watch with a second hand
- Two note pads (one for each Monitor and Tester to document incidents or disturbances that may affect test results).

Before the Students Arrive at the Testing Location

The seating in the testing location should be surveyed to make sure there are enough chairs or desks. Seating of students should be planned so that the Tester can be clearly seen but borrowing of answers will be minimized.

If it is necessary to gather and escort students to the testing room, the Monitor should review the Test Roster (or Test List) for location of students by teacher's name, and identify the teacher's location on the school map. If the students come from more than one classroom, the Tester will assist the Monitor in collecting and escorting students.

Ten minutes before the test session is to begin, the students should be brought to the testing room.

As Students Arrive and Before Testing Begins

Students should be assigned seats as they enter, according to the seating plan.

The Tester will introduce himself and the Monitor and will explain briefly the purpose of testing and the schedule of sittings for the particular group. The purpose of the testing can be explained:

"As you know, you have been involved in a special program this year and we're interested in knowing just how much it's helped each of you. One way to find out is by testing and that's the reason you're here today. The tests you'll take are very important and I know that each of you will do your very best."

If students know what is expected of them, they will be more able and willing to do their best. SRI will then, in turn, be able to obtain accurate estimates of pupil achievement.

After introductory remarks, the Monitor will distribute the test booklets and pencils by calling out the name of each student to ensure that each student receives the proper test booklet. As the student receives his test booklet, the Tester will verify, on the Test Roster (or Test List) the ethnicity and sex of the student noted, or fill in this information if it is missing. The Tester will put a check in the appropriate sitting column if the student is present. If the student is absent, the Tester will put a check in the "See Tester Log" column. The Monitor will then band together any undistributed booklets with the Test Roster (or Test List) and Tester Log. These will be set aside and testing may begin.

Testing Procedures

The following procedures must be followed to maintain an effective testing environment and provide uniformity of procedures among test groups.

Maintaining Control

The Tester should assume control of the group from the beginning and, at the same time, make every effort to maintain the confidence level of the students. The students will be reminded at the beginning of each test that they are not expected to get every item right but that they should do the best they can.

Once the test has started, all remarks should be grouped directed such as:

"Let's all do our own work."
"Let's all work quietly."
"Let's do the best we can."
"We're all working very well, etc."

Remarks should always be in the form of a directive, never in the form of a question.

Practice Items

Practice items are provided on all tests, except the IAR and Student Attitude Questionnaire, to insure that the students know how to mark their answers. As the students do the practice items, the Tester and Monitor should check to see that marks (X or blackened oval under, or next to, test items) are discernible, that they appear in the space provided, that answer selection is being completed quickly, and that only one selection is made for each practice item.

If clarification is necessary, the Tester may demonstrate on the chalkboard how the students should mark their answers and may repeat the practice item questions.

Reading Test Directions

Once the test has been started, the Tester must read the test directions exactly as they appear in the Examiner's copy. The Tester should never elaborate on the directions or provide his own interpretation. Neither the Texter nor the Monitor should provide a clue as to a correct or incorrect response in any manner (e.g., tone of voice, facial expression, etc.). If the Tester judges that most students in the test group did not hear or understand an item, the item may be repeated.

Pacing

On timed tests, students work at their own pace. If all students finish before the allowed time has expired, the test may be terminated. However, the full amout of time allocated for the sitting must be provided any student who wishes to use the remaining time to work on the test. Students who finish early should be encouraged to remai quiet so that those still working can complete the test undisturbed.

On untimed tests, the Tester should move the students along at a pace rapid enough to maintain their attention over the duration of the testing period (i.e., allowing just enough time for the students to mark their answers, but not enough time to look ahead or back to previous answers).

Monitoring

The Monitor oversees the student testing activity to make sure that test results accurately reflect the capability of the individual student to respond. (The Tester will also assume the role of Monitor when not reading instructions to the students.)

Monitoring is a very important part of testing. Possible problem situations and suggested reactions to those situations will be detailed below but, the general requirements for good monitoring can be summed up as follows:

- Be alert.
- Keep moving within the testing room.
- Do not help students with answers, but know ahead of time the page they are to be working on.
- Know ahead of time the common testing problems that can occur.
- Be willing to act immediately to remedy a problem situation.
- Use your pad and pencil to document those problems.

Always be alert. This is the key to effective monitoring. Your eyes should always be on the move, watching for problem situations. In addition to your eyes, you too should be constantly on the move within your designated area. (If the Tester is also monitoring, he should be responsible for half the room and the Monitor the other half. This should be determined before testing begins.) Never stand behind or beside a student and watch him work. You may pause a moment to check for problems, but move on quickly.

Never help students with answers but know ahead of time what he's supposed to do. Listen carefully to the Tester so that you will know what pages the student must complete. If a student appears to have finished his test, always check his booklet to make sure all pages have been completed.

Study the list of possible problem situations and prepare yourself to respond in the most positive and appropriate manner. Never single out students for praise or to express displeasure, but respond when necessary to remedy any situation which will negatively affect test results. Always minimize contact with students who want excessive attention.

Always document on your note pad any situations which require your attention. This documentation will be needed to explain any test invalidations which will be recorded on the Tester Log following completion of the test sitting. Typical situations which might require test invalidation <u>if not corrected</u> are: a student borrowing answers from his neighbor; a student marking responses in an incorrect manner (marking multiple answers, marking the answers outside the designated area); a student working on the wrong subtest; a student not finishing a subtest, thinking he's through but having more pages to complete; or exceptional classroom disturbances.

There is no way to anticipate all the problems that might arise during a testing situation. However, there are certain guidelines which can make monitoring easier and more effective. Following are some possible problem situations and suggested responses:

1) Several students seem confused when the Tester is reading instructions.

   Monitor: Get the Tester's attention and quietly indicate the need to repeat instructions.

2) During an untimed test, students seem either restless or too rushed.

Monitor: Get the Tester's attention and quietly indicate the need to "go faster" or "slow down a little."

3) Student is working on the wrong subtest or on the wrong page, or turns more than one page when asked to "turn the page."

Monitor: Turn to the correct page in the booklet and place it in front of the student.

4) Student is observed borrowing answers from his neighbor.

Monitor: Act immediately. Lightly place a hand on the student's shoulder and turn him back to his test. The Tester should direct a statement to the entire class -- "Let's all do our own work." If the student persists, quietly lead him by the hand to another seat, if available. (Do not, however, place the student outside the test group, i.e., in the corner of the room.) If moving the student is impossible and if borrowing persists, the test must be invalidated. Document the behavior and the student's name on your note pad. Arrange for different seating for the student on the next subtest or during the next sitting.

5) A student is not marking responses in a correct manner; i.e., marking multiple responses to one question.

Monitor: Move your hand across the range of choices and say "you have several choices but select only <u>one</u> answer for each question." (If the student persists, the test must be invalidated. Note the behavior and student's name on your note pad.)

i.e., marking answers in the wrong place.

Monitor: Move your hand across the range of choices and make an appropriate comment such as "Be sure you fill in the answer where it belongs," "Fill in the oval next to (or under) the answer you've selected," or "Put the X on the one choice you've selected," etc.

i.e., marking only one answer when there are several questions connected with a story and there is a choice of answers for each.

Monitor: Draw your hand across the range of choices for the questions left unanswered and quietly say "Be sure you answer all the questions to each story, if you can," or "Notice there is more than one question to each story."

6) A student appears to be finished with the test but is, in fact, not finished.

Monitor: Whenever a student appears to have finished a subtest, always check the subtest to make sure all pages have been completed. If the test is unfinished, turn to the unfinished section and draw your hand across the portions still to be completed. The Tester may say to the class "Be sure to do all the pages, to the bottom (or middle) of page __ where it says STOP."

7) There is an exceptional classroom disturbance: i.e., two students start fighting.

Monitor: Remove both students and their test booklets from the testing area. Try to get them to exercise self-control so they can be returned to the test area. If they are returned, place them at opposite ends of the room. Note on your pad the incident and their names. Check the clock and note the time the students were removed and then the time returned. If the disturbance occurs during an untimed test, note the test item each student was working on at the time of his removal and the test item being administered at the time of his return.

If students cannot be returned to the testing area, take them to their regular classroom teacher(s) or the principal's office if the class is not in session at its regular location. Notify a staff member there of the circumstances to ensure the students will be supervised until they can be returned to their regular classroom teacher.

i.e., a student asks if he can go the the bathroom.

Monitor: Quietly say, "We're almost finished, I'm sure you can wait."

i.e., students finish their tests and become restless.

Monitor: Close the student's test booklet and ask him to work on the front cover. Or, turn the booklet over and quietly suggest the student "draw something" on the back.

NOTE: A student should be removed from the testing group only if he becomes ill or his behavior is so disruptive that it is disturbing the rest of the group.

8) Students seek attention from the Monitor.

i.e., student asks if his answer is correct.

Monitor: Quietly say "Just do the best you can."

i.e., the student smiles at the Monitor each time the Monitor passes by.

Monitor: May smile in return and move on past the student. Try to make an effort to minimize contact with the student during the remainder of the sitting.

This is not an exhaustive list of possible occurrances, but most incidents fall into the general categories described above. Responses suggest appropriate monitoring behaviors.

Collecting Test Booklets

After all students have finished their tests or the Tester indicates the session is over, the Monitor (and Tester, if the group is large) will collect the test booklets (within their predesignated areas) by moving up one aisle and down the other until a test booklet (and pencil, if the sitting has been completed) has been collected from each student.

If a second test booklet is to be distributed during the same sitting, the students may stand and stretch for a few moments before proceeding. Again, the Monitor should call out the names of the students to ensure that each student gets the correct booklet. After the last test has been completed, the booklets are to be collected as described above.

Students should be escorted to their classrooms and reminded to pass quietly through the halls if other classes are in session.

Filling out the Tester Log After Students Have Left the Test Location

The Tester Log must now be filled out for the sitting just completed. There should be a Tester Log for each test group, just as there is a Test Roster (or Test List Worksheet) for each. The top portion of the Tester Log will have been filled out by the Site Assistant. The information will correspond to that on the Test Roster (or Test List worksheet).

The remaining portion of the Tester Log is divided into three sections--"Rating of Group Test Conditions," "Invalidation of All Subtests in a Group," and "Invalidation of Individual Student Subtests." (See Exhibit 4.)

Following the completion of each test sitting, the Tester and Monitor should discuss any incidents they have recorded on their note pads to determine if any were serious enough to have affected the performance of a student or group of students. The Tester should then fill out the Tester Log.

Rating of Group Test Conditions

The Tester should rate the test conditions for the group as a whole by entering a check (✓) on a scale of excellent to poor. Factors to be considered in rating should include comfort of the testing location (heat, lighting, facilities); outside distractions; cooperation of students, and so on.

Invalidation of All Subtests in a Group

The subtests of all students within a group can be invalidated, but only if testing conditions were so poor that the entire group was penalized. For example, if a fire occurs during a timed test and school is interrupted or dismissed, the Tester should record the time of interruption on the Tester Log and have the students close their booklets immediately. The booklets should be collected and an effort made to reschedule the remaining time allotted. If this is impossible, then the subtest must be invalidated for all members of the group.

The same procedures should be followed in the case of an interruption of an untimed test that requires the students to leave the classroom. In this case, however, the item being administered should be noted on the Tester Log and an effort made to reschedule the remaining time.

Exhibit 4

TESTER LOG

PROJECT INFORMATION PACKAGES EVALUATION
Spring 1976

PROJECT NO. ____________ PROJECT NAME ______________________________ CITY ______________

SCHOOL NO. ____________ SCHOOL NAME ______________________________ GRADE ______________

I.D. NUMBERS ON TEST LIST: From ____________ to ____________

The purpose of the Tester Log is to (1) give a general rating of test conditions as they apply to the group being tested, (2) provide means for invalidating all subtests within a group, and (3) provide a record of individual subtests that are invalidated.

1. Rating of Group Test Conditions: In the space provided below rate group test conditions for each sitting.

	Excellent	Good	Fair	Poor	Comments
Sitting 1					
Sitting 2					
Sitting 3					

2. Invalidation of All Subtest in a Group: Invalidation at the group level can occur only if testing conditions were so distracting that the entire group was penalized. Invalidation notations should be made in the comments section.

3. Invalidation of Individual Student Subtests: The decision to invalidate a subtest will be made by the test administrator, in accordance with instructions provided in the test manual. Invalidation codes and conditions are:

1 -- Student refuses to respond throughout most of the subtest

2 -- Student borrows answers consistently

-- Student marks multiple answers consistently

4 -- Student becomes ill during the subtest

5 -- Student was absent

6 -- Student worked the wrong subtest

7 -- Student is in special education

8 -- Student has a severe physical/mental handicap

9 -- Other - specify in the comments section

The spaces below are to be used only if a student's subtest is to be invalidated. Record the student's name, I.D. number, name of the subtest and check the code that explains the reason for invalidating the subtest. Write any additional comments desired.

	Reason for Invalidation			Comments
NAME ______________________	1	2	3	
I.D. # ______________________	4	5	6	
SUBTEST ______________________	7	8	9	

CONTINUED ON THE BACK SIDE

In either case, complete documentation should be made in the Comments section of the Tester Log. If space is insufficient, additional pages should be attached.

Invalidation of Individual Student Subtests

Individual student subtests may be invalidated by the Tester if one or more of the nine specified conditions exist. The specified conditions and associated code numbers are as follows:

Code 1	Student refuses to respond throughout most of the test.
Code 2	Student borrows answers consistently.
Code 3	Student marks multiple answers consistently.
Code 4	Student becomes ill during subtest.
Code 5	Student was absent.
Code 6	Student worked wrong subtest.
Code 7	Student is in special education.
Code 8	Student has severe mental/physical handicap.
Code 9	Other - specify in Comments section.

Codes 1, 2, and 3 should be considered only after the Tester or Monitor has tried several times, with no success, to get the student to respond, to stop borrowing answers, or to stop marking multiple answers to a single question. Code 6 should be considered only if the student was not caught in time to start on the correct subtest. It should not be necessary to use codes 7 or 8 because those students should have been screened out of the test sample already. They are there to be used only in case of a screening error. Code 9 allows the Tester to invalidate a subtest for unforseen reasons not covered by the other codes.

Code 9 should be used in case a student's test in invalidated because of a disturbance. Do not use Code 1 for such cases. Code 1, refusing to respond, is not the same as causing a disturbance.

If a student must leave the sitting for a doctor's appointment, for instance, then Code 5 should be used for any subtests he missed. Do not use Code 9.

If a student is late to class and the test administration has begun, the student may be allowed to begin the test, aided by the Tester or Monitor to find the proper starting place. The word LATE and the name of the subtest being taken at that time must be written on the front cover of the student's test booklet. The incident must be documented on a note pad also.

There are two kinds of tests--the timed test which is <u>not</u> orally administered, and the untimed test which <u>is</u> orally administered. For the timed tests, after the booklets have been collected, the LATE booklet should be checked to see if the subtest which was started late was completed. If it was <u>not</u> completed, the subtest must be invalidated and coded 9. If it was completed, the LATE notation remains but no invalidation code should be entered.

For the untimed test, there is no way a late start can produce valid test results. On the untimed, orally administered test, the student may be allowed to begin late simply to keep him occupied during the test period, but the test must <u>always</u> be invalidated.

Any time a student's subtest is invalidated the student should be allowed to remain in the test location until the testing is completed, unless he is disrupting others around him.

Steps to follow when invalidating subtests:

1) You first observe a disturbance or unusual behavior which may affect test results.
2) You attempt to correct the situation.
3) If the situation cannot be corrected, you describe the occurrance briefly on your note pad, check the student's booklet cover and note his name also.

4) The decision to invalidate is made by the Tester after the test sitting is over. The Tester and Monitor discuss the occurrance and review their notes before the decision is made.

5) The invalidation must then be completely documented.

   a) The name(s) of the subtest(s) invalidated should be entered on the front cover of the student's test booklet.
   b) The proper entries should be made on the Tester Log according to instructions below.
   c) A check should be put in the column "See Tester Log" on the Test Roster (or Test List worksheet). This will alert coders at SRI that there is an invalidation or unusual circumstance regarding that student's subtest(s).

Whenever a student's subtest is invalidated, the following information must be entered in the boxes provided on the Tester Log.

1) Student's name.
2) Student's three-digit ID number.
3) Name or names of subtest(s) invalidated during the particular sitting.
4) Reason for invalidation checked (Codes 1-9).
5) Reason for invalidation provided, briefly, in the Comments section for any invalications other than for absence (Code 5).

A student's name may appear on the Tester Log more than once--up to as many times as there are sittings for his test group.

The <u>name</u> of <u>each</u> subtest invalidated during a sitting must be entered. You should not note "sitting 1" for instance, rather you should note the names of the actual subtests invalidated during sitting 1. You should not note just "math" for instance, but should note which math subtests were invalidated, i.e., "Math Concepts," "Math Comprehension," or "Problem Solving." You should also specify whether "Word Knowledge" or "Reading" has been invalidated, or enter both names, if both were invalidated.

SRI must always know the reason a subtest has been invalidated or has been left blank. A brief explanation <u>must</u> be entered in the Comments section <u>each</u> time an invalidation occurs, except for absence.

In all cases where Test List worksheets have been prepared from the Test Roster, please transfer the following information from the worksheets to the Roster after all sittings have been administered to the group appearing on the Test List worksheet. Transfer <u>all check marks</u> in the columns headed "Sittings," and "See Tester Log." <u>All Information in the Comments sections</u> should also be transferred to the Test Roster from the Test List worksheets. Remember--do not discard the Test List worksheets. They are to be banded together with the completed test booklets and the Tester Log.

Section IV

SITE ASSISTANT'S INSTRUCTIONS FOR RETURNING TEST MATERIALS TO SRI

Before shipping tests back to SRI, the Site Assistant must check the booklets against Test Rosters (or Test Lists) to make sure that all used test booklets have been accounted for. Do not change marks made by children.

Packing and Shipping Booklets

When testing has been completed for a test group, the test booklets and the appropriate Test Roster (or Test List) and Tester Log should be grouped together, rubber banded, and placed in the test carton. Place as many groups' tests in a carton as will fit. Tests from the same group should not be split into separate test cartons.

All unused test booklets, both labeled and unlabeled, should be returned to SRI. The number of unused booklets should be noted in column 5 of the Packing Invoice. Column 4 and 6 of the Packing Invoice should also be completed by the Site Assistant.

When cartons have been packed for return, they should contain:

1. Unlabeled test booklets.
2. Labeled but unused test booklets.
3. Test booklets used by the students. These booklets should be grouped with their Test Rosters (or Test List worksheets) and Tester Logs.
4. All other miscellaneous test materials that were originally included in the carton(s) EXCEPT THE PRECUT SEALING TAPE AND RETURN ADDRESS LABEL(S).
5. The Packing Invoice should be placed on top of the contents of one of the cartons. In those sites where Test List worksheets were prepared from Test Rosters, the original Test Rosters should be packed with the Packing Invoice.

The test materials should be shipped to SRI via United Parcel Service. UPS will not send materials COD, therefore Site Assistants should be prepared to pay for the shipping costs and then include the cost on expense invoices to be reimbursed. The test materials should be shipped as soon as possible after testing has been completed. The Site Assistant should obtain a shipping number and estimated time of departure. The Site Assistant should then call Ben Samson collect at (415) 326-6200, ext. 3118, and report the shipping number, number of cartons being shipped, and estimated time of departure.

Appendix B

CONVERTING STANDARD SCORES TO PERCENTILE RANKS AND DETERMINING THE EXPECTED SPRING SCORE FOR THE NORM-REFERENCED ANALYSIS

Appendix B

CONVERTING STANDARD SCORES TO PERCENTILE RANKS AND DETERMINING THE EXPECTED SPRING SCORE FOR THE NORM-REFERENCED ANALYSIS

Percentile ranks are obtained from standard scores by using Table 3 in the relevant MAT Teacher's Handbook. Table 3 has beginning-of-year norms for fall testing and end-of-year norms for spring testing at each grade level for which the tests are appropriate. If a particular standard score does not appear in the table, the handbook instructs the teacher to use the percentile rank corresponding to the next higher standard score.

We determined this "rounding-up" procedure to be too insensitive for the norm-referenced analysis. An interpolation method was used to provide more accurate conversion of PIP and site level mean standard scores to percentile ranks. The method, which takes into consideration the underlying normality of the standard scores, is as follows:

(1) Find the two percentile ranks corresponding to next higher and next lower standard scores in the table.

(2) Look up z scores corresponding to the upper and lower percentiles in a table of the standard normal cumulative distribution function.

(3) Perform a simple linear interpolation of the z scores to determine the z score corresponding to the particular standard score.

(4) Look up this z score in the normal distribution table to determine the interpolated percentile rank.

The expected spring standard score for a given fall standard score is determined by a simple linear interpolation between the beginning-of-year and end-of-year standard scores. For example, if the fall mean standard score of 59.3 lies between standard scores of 59 and 62 in the beginning-of-year table and if these two standard scores correspond to standard scores of 64 and 66, respectively, in the end-of year table, then the expected spring standard score becomes $64 + (66 - 64) \times (59.3 - 59)/(62 - 59)$, or 64.2.

Appendix C

CLARIFICATION OF PIP SPECIFICATIONS
RESULTING FROM THF WASHINGTON CONFERENCE

Catch-Up Group Meeting
Washington, D.C.
18 September 1975

PROJECT CATCH-UP

Summary of Revisions and Clarifications

Public Relations

- Daily contact between project teachers and regular teachers is vital. The teachers' lounge and the lunchroom are excellent places for such contact, as is the playground.
- One of the first faculty meetings of the year can be held in the Catch-Up lab.
- Parental involvement is more difficult to bring about in some places than in others; less frequent involvement is expected if parents live a great distance from the school than if they live next door. Potluck dinners once or twice a year are one way of getting parents involved.

Staff

- Teachers in Catch-Up should work four hours a day, not three, in order to work with just two, three, or four students at a time. They are generally paid on an hourly basis.
- Part-time staff is essential to the project.
- Aides do the complete job of instructors.
- Teachers and aides should always maintain a positive, success-oriented approach in working with their students. Some suggestions are to provide a badly behaved student with an excuse ("You're too tired today. Why don't you come back tomorrow when you've had more rest.") and to encourage a poor reader to read simple books by saying he or she may someday be a father or mother and will want to read baby books to the children.

- Teachers and aides base their individual responsibility for the gains of 18 or 10 students on median, not mean, averages.

Materials

- The PIP listed eight years' accumulation of materials. All of this, of course, need not be acquired by sites in the first year of operation.
- Criterion-referenced tests may be used as teaching tools, this is usually done about once a week. These tests are a good means of keeping teachers on target in relation to individual students' needs.
- The danger in using Catch-Up materials in the regular classroom is that they will lose their special status and may bore the students.
- If materials are used that correlate with regular classroom materials, the child's confusion is reduced.
- Random House math materials might improve Catch-Up math instruction.
- Catch-Up is not the type of lab in which children move from one spot to another and in which materials and instructor are permanently stationed. Rather, it is a place where materials, instructors, and children all move about freely.
- In scheduling, it works well for each teacher to have access to a particular machine or teaching tool on one day of the week. The teacher can decide whether to use the machine that day and, if not, can give another teacher permission to use it.

Other

- Parent aides are used primarily to help out in bilingual labs. Their use in regular Catch-Up labs is limited.
- In some ways, the original Catch-Up design is geared toward bilingual Spanish-speaking children. This design can be adapted to match new contexts; one site, for example, introduced materials for black awareness.

Conquest Group Meeting
Washington, D.C.
18 September 1975

PROJECT CONQUEST

Summary of Revisions and Clarifications

Public Relations

Positive relations between project and nonproject staff are crucial in maintaining the necessary district support for the project.

- The attitude of project staff should be that they are there to assist regular classroom teachers. Scheduling is to be worked out as much as possible to the convenience of classroom teachers. In cases where a mutually agreeable schedule cannot be arranged, the child can be placed on a waiting list or in the control group.
- Friday afternoons can occasionally be used to allow project teachers to observe regular classrooms and meet with teachers. Clinicians should take students' Deficiency Checklists along to discuss with teachers.
- Teachers report to principals at beginning and end of year and to parents three times a year (beginning, middle, and end).

The importance of involving parents as well as nonproject staff was stressed. Various techniques have been found successful:

- Invite parents, teachers, and administrators to chili suppers; "marathon meetings," where breakfast, lunch, and dinner are served; awards luncheon at the end of program.
- Babysitting and transportation services should be provided for parents.

Staff Issues

The major issue was the role of the supervising clinician.

- The main duties of the supervising clinician are (1) to assist the project director with training and administration, (2) to monitor and assist clinicians, and (3) to teach his/her own students.

- The amount of time spent on any one of these duties depends primarily on the number of centers. At sites where the project is relatively small (e.g., fewer than 10 centers), supervising clinicians will have their own students to teach and will spend one morning or one day per week monitoring and assisting clinicians. At sites with a large number of centers (e.g., over 20), supervising clinicians may spend the major portion of the week observing clinicians and have no students assigned to them.
- Calling the supervising clinician a Reading Coordinator or a Consultant and clarifying the role as one of assistant rather than supervisor will help avoid problems with staff resentment and union regulations.

Training

- Length of preservice training for the first year should be two weeks. For every year thereafter, new teachers should receive two weeks' training, and teachers continuing in the project should receive at least one week. Preservice training should cover diagnosis only. Materials and remediation techniques should be taught just prior to remediation and during in-service sessions.
- Each clinician is to administer, take, and interpret each test during preservice training.
- During training, clinicians should work through a sample case from diagnosis up through remediation.

Instruction

- Diagnosis
    - The following 12 diagnostic instruments are to be used. The importance of administering the instruments in this order was stressed by the originating project director. Tests that do not appear on this list have been deleted from the battery.

General Information Sheet

Teacher Referral

Health Screening--Audiometer & Titmus (to be administered by nurse)

Slosson Intelligence Test (SIT)*

Slosson Oral Reading Test (SORT)

California Ach. Test† (reading rooms); Gates-MacGinitie† (clinics)

Bond-Balow-Hoyt (reading rooms); Stanford Diagnostic Test (clinics)

Spelling Inventory--Betts (reading rooms); Kottmeyer (clinics)

Reading Inventory--Informal (reading rooms); Subjective (clinics)

Programmed Reading Placement Test*

Interest Inventory

Deficiency Checklist.

- It was agreed that it is the <u>process</u> of diagnosis that is important rather than the specific tests used. The idea is to get the information provided by these instruments while administering the smallest number of tests possible.
- Replicating sites have found certain other tests useful (e.g., the Wepman or Peabody for visual/auditory discrimination; Fountain Valley for word recognition).

- Individualization
  - Remediation proceeds from the Deficiency Checklist. Teachers should find comprehension passage appropriate to each deficiency, then work backwards to vocabulary and phonics.

---

*Results from as much as two years previous may be used.

†If a score is available from a test providing national norms on vocabulary and comprehension, this test is not needed.

- Instruction should be adapted to student's interests, as revealed by Interest Inventory.
- Individualization depends on detailed record keeping.

- Instructional facilities
  - Principals should be involved in designation and allocation of facilities before start-up of program.
  - The reading room/clinic combination should be in one room; separate centers should be housed in separate rooms or portable facilities.
- Instructional materials
  - Regular district basal series is not to be used.
  - Programmed reading series should be Sullivan, unless Sullivan is being used in the regular classroom. If this is the case, another programmed series covering comprehension may be substituted. It is important that one continuing series be used daily.
  - To reinforce skills, move from Conquest to Dr. Spello.
  - Conquest workbook is not to be used as a consumable.
  - When there is a suggested order in materials (e.g., the Phonovisual), follow the sequence described.
  - The guideline is 10 minutes on each activity, but this is not rigid. At the beginning of each period, 5-10 minutes should be spent establishing rapport with the children.

Record Keeping

- To achieve consistency, use one system for primary and one for intermediate. These systems are to be worked out by the teachers.
- Clinicians should take three minutes at the end of each period to record what children have done.

Reading Rooms Versus Reading Clinics

- Structure and format are the same.
- Materials differ according to level.

HIT Group Meeting
Washington, D.C.
18 September 1975

PROJECT HIGH INTENSITY TUTORING (HIT)

Summary of Revisions and Clarifications

Public Relations

Good public relations lead to long-term survival of the project; therefore,

- The project director must ensure that principals are consulted in matters that affect their schools as a basis for winning their long-term support. The principals should gradually gain a sense of ownership of the project.
- Project teachers should ensure that their colleagues support the project. Support is more likely when classroom teachers are involved in the selection and scheduling of tutors and tutees and when project teachers maintain daily contact with the instructional staff. If project teachers are new to the school, winning respect requires extra effort.

Tutors/Tutees

- Tutors should mainly be eighth graders and tutees sixth graders. The age difference was found to be important at the originating site.
- The maximum number of tutees to serve per period in HIT is 12. More tutoring pairs are hard to monitor effectively and paperwork is excessive. Since the program is designed to benefit tutors as well as tutees, this still allows 24 students to be served in each period.
- The tutor pool should be about twice as large as the number of tutees. Tutors usually come three times a week; none come only once. They come at different times so they do not miss the same class more than once a week. Tutors are never taken out of reading

and math classes. Tutors make up homework for classes they miss. It becomes easier to recruit tutors if administrators adopt a policy that they may miss a class to come.

- Tutees do not make up homework in classes missed; the cooperation of the classroom teacher is essential.

Instruction

- Each center should have only five half-hour tutoring sessions per day. Longer sessions should be avoided if possible, since it is difficult to maintain the intensity level characteristic of HIT for more than half an hour.
- Materials should be on student desks when they arrive for tutoring. Tutoring should begin right away without any distractions from the teacher, such as roll call. Absences can be noted from unused student folders. Materials should be placed where tutors can pick them up as needed without the teacher's or aides' help. Tutoring should be going on during the entire period with no time allocated to pep talks, discussions of discipline, or delay for passing out materials.
- Record keeping and setting out folders and materials should occur during breaks between each session. Teachers and both aides should be walking through the room listening to tutoring and helping as necessary while tutoring is going on. Teachers and aides occasionally tutor to help students having special problems. Enough tutors should be recruited so that teachers do not perform this role merely because there are not enough tutors.
- No written answers or drill should be assigned in HIT reading centers. This would slow down oral reading practice and be tedious to students.
- Avoid lengthy explanations that take time students could use in active skills practice.
- Students may jump several pages in Sullivan if they get over 94% correct for three days. If they pass a section test in the middle (or at the end) of a book, they may jump to the next section (or book). At the beginning of the year they may jump through many

books in a few days if they did poorly on the placement test but are able to do the work.

- When a tutee, such as a hyperactive child, does not respond well to programmed and drill materials, the teacher may work with him to diagnose his or her reading or math problems and assign other types of work. It helps to assign very patient and skilled tutors to work with such students.
- Teachers should not be concerned about tutees "peeking" at answers; they should be getting 90%-94% correct and can learn by looking up the others on the answer sheet.
- Never publicly correct students, especially tutors, for misbehavior during a session. Talk to tutors after class in a private room, alone. If they do not want to be tutors, replace them. Make tutoring desirable by treating tutors as paraprofessionals. Emphasize that they are teachers, too, but will not have permanent positions if they do not accpet the responsibility. In the beginning of the year, nominate them as "chosen" rather than asking for volunteers.
- When tutees make an error, an "H" for "help" may be placed on the tally sheet instead of a zero, which has more negative connotations for students.

Materials

- If older or younger students attend HIT, materials other than Sullivan may be needed for them. Materials should be selected that lend themselves to fast-paced tutoring. They should have:
  - Simple directions or a repetitive layout that makes directions unnecessary to repeat.
  - Answers easily found and read by tutors (not in tiny print in a difficult-to-find teacher's section).
  - Unambiguous answers (no subjective questions with several possible answers for tutors to judge).
- Drill in math uses many different materials designed to teach basic facts. These may be commercial or teacher-made. Records on which sets of facts each

student has mastered are kept in the student folder. Teachers should devise sufficient drill materials (e.g., six or seven ways of teaching multiplication tables, and exercises such as measuring the room to learn the metric system) to keep drill from being unnecessarily monotonous. Points are given for drill partly so that the error rate in drill is controlled at 90%-94% correct.

- Drill materials used in reading can include word lists derived from lists in Conquests in Reading, Why Johnny Can't Read, or other sources. Drill in reading is aimed at fluent word recognition. Spelling or writing words letter by letter interrupts this fluency and should not be done. Students should work with a set of words until they can read 90%-94% of them fluently without stopping to decode them. Variation can be added to drill by occasionally making a game or contest out of drill words.
- Occasionally for variety in reading, auxiliary materials such as plays may be used in place of programmed readers.

Rewards

- Attractive certificates printed locally can be given to tutors at the end of the year, as well as awards for the highest achieving boy and girl tutee and tutor. Tutors can also be given holiday presents, such as small wallets or bracelets. Only tutors who have come at least a minimum number of times should come on field trips. It is helpful to consult with tutors regarding the types of rewards they would like (e.g., where to go on field trips).
- One point is awarded for each drill item and one point for each problem in math and sentence in reading. If too many points are earned for the reward budget, the "price" of items (in points) may be changed.
- Tutee rewards may include fruit, model airplanes, cafeteria passes, or other items, but attractive candy, such as chocolate bars, should be among the options offered students each time rewards are given. Teachers should decide how to spend money for rewards.

- Tutors may receive 200 points for each session they attend. They use these points toward candy rewards This is optional.

Other

- New students may join the program mid-year if there is room. They begin at their level in Sullivan.
- Students who complete all the programmed materials may drop out of the formal program mid-year, but the teacher should continue to see these students frequently, invite them to visit, and ensure that they feel they are receiving special attention even though they have finished with Sullivan.
- Be wary of including special education (retarded) students in HIT if these students are not integrated into all classes. Many HIT students may be sensitive to being labeled slow learners. The project is not designed for special education students.

IRIT Group Meeting
Washington, D.C.
18 September 1975

PROJECT IRIT

Summary of Revisions and Clarifications

General

- Good relations with the sending school principals and teachers and with district administrators are critical to project success. Maintaining good relations is an important part of the jobs of IRIT project directors and teachers.
- IRIT should make an effort to help regular teachers improve their teaching of reading. Intern programs, demonstrations, and inservice sessions were described.

Project Organization

- Where practicable, drawing all 45 students in a cycle from one school is preferred--both by sending teachers and for logistics.
- When all IRIT students from a single classroom attend the same cycle, the regular class size is significantly reduced. This has proved to be one of the major attractions of IRIT and should be tried in new sites if at all possible.

Student Selection Process

- Before each cycle, the entire team should meet with sending teachers to explain the project and start the selection process. This meeting is a critical part of establishing good relations.
- Teachers are asked to nominate about 60 students. The IRIT team selects 45.
- No student should start more than two weeks late. Students dropping out after this time should not be replaced.

- The first cycle should start about two weeks after the beginning of school in the fall. This allows time for the selection process as well as some training and room preparation.
- The last cycle should end as near to the end of the regular sessions as possible to minimize disruptions to sending rooms at year's end.

Instruction

- Coordination of the three reading areas requires daily meetings of the team. Decoding can be used as the core for discussing each student.
- Individualized reading requires an exceptionally energetic, personable, and creative teacher.
- IRIT teachers generally specialize in one area rather than rotating from area to area within a given year, but may change from year to year.
- IRIT trains its own substitute teachers so that they will be familiar with IRIT procedures.

Materials/Equipment

- Basal readers should be integrated into the curriculum if the regular teachers so request. They should be incorporated into the IRIT approach, but not to the exclusion of other core IRIT materials.
- Whether or not the basal reader is used in IRIT, the regular teacher needs to know where to place each student after the cycle. IRIT teachers should give an appropriate placement test at the end of each cycle and advise the sending teacher.
- IRIT does not place excessive emphasis on the use of teaching machines, although they can be useful for practice and as motivators.

PTR Group Meeting
Washington, D.C.
18 September 1975

PROJECT PROGRAMED TUTORIAL READING (PTR)

Summary of Revisions and Clarifications

Seven topics were discussed by the participants in the PTR group at the replication conference. They are reported in their order of importance to the discussants, as perceived by the recorder.

Teacher Participation in Selection of Students for PTR

- Two objections were raised to including teachers heavily in the student selection process. First, it was noted that including teachers' subjective opinions about students as a basis for selection could (in Dallas, would) corrupt the local evaluation design. Second, it was felt that asking for teachers' judgments for selecting students was an inappropriate way to elicit their approval for the project. The way to select students, it was agreed, is to use test scores. Teachers who strongly disagreed with the selections made this way could review test results and discuss their opinions with the director or supervisor. If a strong case was made, arrangements could be made to retest the student in question. This was the procedure used in Farmington when PTR was validated.
- Recent refinements to the student selection procedure used in Farmington were discussed. They now use a combination of test scores and teacher rating (without teacher review of the test results before the rating). This method was considered unfeasible for other locations, as Farmington has a computer program to compile the raw data and a rather complicated formula for incorporating teacher rating with test scores. Farmington now uses spring posttest scores in selecting students rather than administering a pretest in the fall. (They tutor the first four grades.) In selection also, the previous year's teacher ratings are used, not the new teacher's ratings.

- In terms of selecting children, something should be said about the populations being served. Farmington is predominantly middle-class and rural. The Dallas population is low income, predominantly black, and inner-city. Canton has a similar population, except it is rural. Farmington reportedly had a number of students "top out" on the pretest (Murphy-Durrell), whereas in Canton a significant number "bottomed out." In Dallas, most were at the low end of the scale.

Teacher Support for PTR

- Discussants agreed that it was virtually impossible to create strong teacher support for PTR prior to its operation for two reasons: First, they cannot yet see the value of the project; second, teachers are inherently threatened by the prospect of others, particularly those not members of the guild, "teaching" their students in a subject as fundamental as reading.
- The best that can be done is to explain PTR as fully as possible to teachers before the tutoring begins, especially pointing out that the tutors do not initiate teaching strategies but are told exactly what to do by the tutoring programs. In Farmington, teachers were brought into four orientation sessions before the project began, on a paid basis. This option may not be economically possible elsewhere. In any event, all three project directors agreed, by the end of the first year of operation the teachers overwhelmingly supported the project because of the results it had achieved, both cognitively and affectively. The replicating sites felt that the only way to elicit teacher support was to demonstrate the project for a year. This meant mandating it and trying to hold the line until teachers realized the positive aspects of the program.

The Alphabet Skills Book

- All three project directors have concluded that The Alphabet Skills Book is indispensable and must be included in the PIP. When tutored students do not know the alphabet skills, the tutors invariably become frustrated and the time is wasted. Both replicating sites obtained copies of The Alphabet Skills Book and found it absolutely mandatory for children without readiness skills.

- The Farmington site has been unsuccessful in its effort to secure permission to incorporate The Alphabet Skills Book into its program. As a result, Farmington is developing its own readiness skills book, which would represent a useful addition to the PIP. The Farmington PTR staff use this readiness skills book exclusively for the first 6-8 weeks at the first grade level to prepare children for the tutoring materials that correspond to basal readers. This also delays introduction of the tutoring kits long enough to minimize the possibility that tutees will move through the basal readers more rapidly than other students in their classrooms.

Training

- The present training is not sufficient. There should be separate training materials for each publisher's tutoring kit.
- A suggested sequence of training steps is: preview, explanation, demonstration, practice. Each task or small group of related tasks should be separate instead of on the same tape sequence. Each step should be more fully explicated. There is not enough practice time allocated in present training as implemented in sites. Farmington had outside training support (University of Indiana) for the first 2-3 days.

Parent Involvement

- Discussion on the issue of getting parents involved revealed that the differences in socioeconomic levels between the originating and the replicating sites were vast. The originating site found it much easier to involve parents, who were typically middle-class. In the depressed areas the replicating sites served, it was almost impossible to get parents involved. Some suggestions for drawing parents out included:
    - Paying them to attend meetings
    - Inviting them to actual tutoring sessions
    - Holding parties, with refreshments, at which their children read

- Giving rewards to parents of high achieving tutees
- Coordinating PTR with the Reading Is Fundamental free book program
- Taking native language tape/slide presentations to parents' homes.

- An encouraging note was that, regardless of parent involvement during the year, there was strong support for PTR among parents, who reported that participating children read a great deal and held positive attitudes toward themselves and learning because of the program. In fact, had it not been for parent enthusiasm, the program would not have survived past the first year in Farmington. Test score gains were nil the first year of the program, according to Dallas Workman. However, the program proved strong in developing positive attitudes among tutored children.

Compatibility of PTR and Nonpublisher Basals

- Dallas Workman reported that many Farmington children who participated in PTR have their basic reading program in basals not connected to PTR-associated programs. Children in these basal programs show no difference in achievement gains from children in the basal reading series that match the PTR program they use (i.e., the Ginn 100 is used in some classrooms, whereas the Ginn 360 tutoring kits are used; some tutored children have their basic reading programs in texts not published by PTR kit publishers, as well).

Technical Assistance

- The three project directors agreed they need to have a resource person whom they can call when confronted with unprecedented project problems in both instruction and management. Replicating site directors were grateful for Dallas Workman's suggestions on the phone. All three agreed to continue to support and be on call for one another and for new replicating sites.

- Dallas Workman felt that, for future PTR replicating projects, two contacts were essential.
  - Someone at the federal level (monitoring agency).

- Someone capable of giving technical advice based on experience (i.e., Farmington, Canton, Dallas).

- Dallas Workman also felt that, if the program was to succeed as a project, it was essential that the projects have the latitude of "adapting" the PIP to local conditions.

R-3 Group Meeting
Washington, D.C.
18 September 1975

PROJECT R-3

Summary of Revisions and Clarifications

Gaming/Simulations/Contracts

- Summer workshops in gaming/simulations should be planned (for cadre teachers) to get a head start on adapting and integrating gaming/simulation into the curriculum.
- Gaming/simulation activities should be incorporated into the reading, math, and social studies curricula as often as possible.
- Contracts are an integral part of R-3. Contracts can be purchased as well as developed by teachers.

Instruction/Materials

- Overuse of one instructional method deadens the effectiveness of the approach.
- Two planning periods a day are needed by all R-3 teachers to plan the curriculum and coordinate among the teams and team members.
- Project director autonomy is essential to Project R-3 success. The project director must have budget control to purchase materials needed by teachers.
- Diagnostic-prescriptive materials should be used extensively in reading and math, although activities should be changed to reduce or eliminate boredom.
- Barbara Evans (Project Director, Lorain, Ohio) agreed to send lists of contract materials she found useful as well as reading materials used for the diagnostic-prescriptive technique.

Public Relations

- Home visitations must be conducted early in the school year. A great deal of effort is needed in obtaining parents' understanding of gaming/simulation and the R-3 philosophy.
- Greater efforts will be made to help other teachers understand the R-3 projects; at the same time, R-3 project teachers will continue to plan together as a group.

Appendix D

INDEPENDENT VARIABLES FOR THE REGRESSION EQUATIONS

Appendix D

INDEPENDENT VARIABLES FOR THE REGRESSION EQUATIONS

The independent variables for computing the regression equation for each grade level for reading and math are indicated by X's on Tables D-1 and D-2. For example, the regression equation for grade 1 reading included seven variables: (1) age, (2) sex, (3) minority status, (4) good-bad rating for Catch-Up teacher responsiveness, (5) well-poor rating for Catch-Up teacher implementation, (6) good-bad rating for PTR (Dallas) teacher responsiveness, and (7) well-poor rating for PTR (Dallas) teacher implementation.

Table D-1

INDEPENDENT VARIABLES FOR THE READING REGRESSION EQUATIONS

Variable	Grade 1	1 (Canton)	2	3	4	5	6	7	8	8 (R-3)
Continuous										
Age	X	X	X	X	X	X	X	X	X	X
Indicator										
Sex (0 = Female, 1 = Male)	X	X	X	X	X	X	X	X	X	X
Minority (0 = Caucasian, 1 = Black or Spanish)	X		X	X	X	X	X	X	X	
Black (0 = Caucasian or Spanish, 1 = Black)										X
White (0 = Black or Spanish, 1 = Caucasian)										X
Well-implemented PIP (see Section 6.2) (0 = Not well implemented, 1 = Well implemented)										
Catch-Up	X				X	X	X			
Conquest	*		X	X	X	X	X			
HIT								X		
IRIT										
PTR	X									
R-3										
Responsive Teachers (see Section 6.2) (0 = Not responsive, 1 = Responsive)										
Catch-Up	X				X	X	X			
Conquest	*		X	X	X	X	X			
HIT								X		
IRIT										
PTR	X	X								
R-3										X
Combination of well-implemented PIP and responsive teachers (see Section 6.2) (0 = Poorly implemented or not responsive, 1 = Well implemented or responsive)										
Catch-Up			X	X						
Conquest										
HIT							X		X	
IRIT				X	X					
PTR										
R-3										

*Insufficient data for the PIP at this grade level.

Table D-2

INDEPENDENT VARIABLES FOR THE MATH REGRESSION EQUATIONS

Variables	Grade 1	1 (Canton)	2	3	4	5	6	7	8	8 (R-3)
Continuous										
Age										X
Indicator										
Sex (0 = Female, 1 = Male)										X
Minority (0 = Caucasian, 1 = Black or Spanish)										
Black (0 = Caucasian or Spanish, 1 = Black)										X
White (0 = Black or Spanish, 1 = Caucasian)										X
Well-implemented PIP (see Section 6.2) (0 = Not well implemented, 1 = Well implemented)										
Catch-Up	X				X	X	X			
Conquest										
HIT							*		*	
IRIT										
PTR										
R-3										
Responsive teachers (see Section 6.2) (0 = Not responsive, 1 = Responsive)										
Catch-Up	X				X	X	X			
Conquest										
HIT							*	X	*	
IRIT										
PTR										
R-3										X
Combination of well-implemented PIP and responsive teachers (see Section 6.2) (0 = Poorly implemented or not responsive, 1 = Well implemented or responsive)										
Catch-Up			X	X						
Conquest										
HIT										
IRIT										
PTR										
R-3										

*Insufficient data for the PIP at this grade level.

Appendix E

MOST FREQUENTLY USED MATERIALS AND SKILLS EMPHASIZED IN CATCH-UP, CONQUEST, HIT AND IRIT PROJECTS, BY GRADE LEVEL

Table E-1

MOST FREQUENTLY USED MATERIALS AND SKILLS EMPHASIZED IN CATCH-UP PROJECTS, BY GRADE LEVEL

a. Grades 1 and 2: Reading

Catch-Up Material	PIP Specified Core or Supplementary Material	Bloomington: Frequency of Use	Bloomington: Percent of Days Sampled	Brookport: Frequency of Use	Brookport: Percent of Days Sampled	Primer: Recognition of Sounds	Primer: Letter Recognition	Primer & Primary I: Decoding	Primary I: Structural Analysis	Primary I--Advanced: Vocabulary	Primary II--Advanced: Antonyms and Synonyms	Primary II--Advanced: Comprehension: Word, Sentence, and Paragraph	Comments
Published reading material													
1. Instant Readers		5	5.7%	0	0%								
2. Library (free reading)		8	9.1	0	0								Enjoyment, word patterns
3. Gateways to Reading Treasures		5	5.7	0	0								Enjoyment
4. Scholastic Individualized Reading	Core	5	5.7	0	0				x	x	x	x	Enjoyment
5. Sounds of Language Readers		4	4.5	0	0								
6. SRA Reading Laboratory:		14	15.9	0	0				x	x		x	Enjoyment, word patterns
My Own Book		4	4.5	0	0	x				x		x	
7. SRA Reading Program	Supp.	24	27.3	0	0					x		x	
8. Systems 80	Core	16	18.2	0	0	x	x	x	x	x			
9. Random House Criterion Reading	Core	12	13.6	58	21.9								
10. First Talking Alphabet	Supp.	0	0	13	4.9	x	x	x		x			Testing in all skills
11. Open House Series--More Power		0	0	18	6.8								
12. Phonics We Use (Old & New)		0	0	44	16.6	x	x	x	x	x			
13. Phonics We Use Learning Games Kit	Supp.	0	0	7	2.6	x	x	x	x	x	x		
14. Specific Skills Series	Supp.	0	0	26	9.8					x		x	
15. Alpha Bingo (game)		0	0	2	0.8	x	x						
16. Dolch Basic Word List		0	0	2	0.8					x			
17. Letter Recognition		0	0	2	0.8	x	x						
18. Sullivan Reading Readiness		0	0	4	1.5	x	x	x					
Other		21	23.9	5	1.9								
Teacher-made reading material		22	25.0	122	46.0								
Total instructional days sampled		88		265									

b. Grades 1 and 2: Math

Catch-Up Material	PIP Specified Core or Supplementary Material	Bloomington: Frequency of Use	Bloomington: Percent of Days Sampled	Brookport: Frequency of Use	Brookport: Percent of Days Sampled
Published math material*					
1. Holt Math Test		2	2.3%	0	0%
2. Sullivan Math	Core	1	1.1	102	38.5
3. Tutor Computor	Core	6	6.8	47	17.7
4. Singer Individualized Math	Supp.	1	1.1	23	8.7
5. Systems 80	Core	0	0	31	11.7
6. Houghton Mifflin Basic Facts & Skills		0	0	11	4.2
7. Houghton Mifflin Skill Sheets		0	0	24	9.1
8. BASE Diagnostic Test		0	0	2	0.8
9. Digitor		0	0	8	3.0
Teacher-made math material*		4	4.5	6	2.3
Total instructional days sampled		88		265	

Note: Frequency tabulations were not made for Providence Forge and Wayne City because we had no data for these grade levels; frequency tabulations were not made for Galax because of inadequate data.

*The skill or skills emphasized in the math materials were not recorded because we were unable to find adequate information on materials used.

Table E-1 (Continued)

c. Grades 3 and 4: Reading

Catch-Up Material	PIP Specified Core or Supplementary Material	Bloomington: Frequency of Use	Bloomington: Percent of Days Sampled	Brookport: Frequency of Use	Brookport: Percent of Days Sampled	Providence Forge: Frequency of Use	Providence Forge: Percent of Days Sampled	Wayne City: Frequency of Use	Wayne City: Percent of Days Sampled	Primer: Recognition of Sounds	Primer: Letter Recognition	Primer and Primary I: Decoding	Primary I: Structural Analysis	Primary I--Advanced: Vocabulary	Primary II--Advanced: Antonyms and Synonyms	Primary II--Advanced: Comprehension: Word, Sentence, and Paragraph	Comments
Published reading material																	
1. Creative Features Structural Analysis		6	3.4%	0	0%	38	11.1%	0	0%				x				
2. Open Court Correlated Language Arts & Reading Program	Supp.	10	5.7	0	0	11	3.2	0	0	x	x	x	x	x	x	x	
3. Scholastic Individualized Reading	Core	4	[illegible].3	0	0	8	2.3	2	0.7				x	x	x	x	
4. Sounds of Language Readers		10	5.7	0	0	0	0	0	0								Enjoyment
5. SRA Reading Program	Supp.	15	8.6	0	0	12	3.5	0	0					x		x	
6. SRA Reading Laboratory:		15	8.6	0	0	30	8.8	0	0				x	x	x	x	
My Own Book		4	2.3	0	0	0	0	0	0	x				x		x	
7. Read story/book (unspecified)		7	4.0	1	0.4	0	0	1	0.3								Enjoyment
8. Random House Criterion Reading	Core	30	17.1	108	38.0	20	5.8	146	48.5								Testing in all skills
9. Systems 80	Core	9	5.1	9	3.2	0	0	4	1.3	x	x	x	x	x			
10. Dolch Word List		0	0	2	0.7	0	0	6	2.0					x			
11. Magic Seasons		0	0	10	3.5	0	0	0	0								Enjoyment
12. Minibike Film & Worksheet		0	0	2	0.7	0	0	0	0							x	Enjoyment
13. Phonics We Use (Old & New)		0	0	55	19.4	0	0	0	0	x	x	x	x	x			
14. Phonics We Use Learning Games Kit	Supp.	0	0	6	2.1	0	0	0	0	x	x	x	x	x	x		
15. First Talking Alphabet	Supp.	0	0	1	0.4	0	0	0	0	x	x	x		x			
16. Specific Skills Series	Supp.	0	0	9	3.2	36	10.5	19	6.3	x				x	x	x	
17. Ginn Elementary English Series		0	0	0	0	5	1.5	0	0								English skills not listed (i.e., verbs & nouns)
18. Words in Motion		0	0	0	0	16	4.7	0	0					x		x	
19. Mission Read		2	1.1	0	0	28	8.2	3	1.0					x		x	Enjoyment
20. Core File (lessons correlated with Random House Criterion Reading)		0	0	0	0	0	0	24	8.0								Materials covering all skills
21. Cyclo Teacher		0	0	0	0	0	0	11	3.7				x	x	x	x	
22. Build A Sentence Game		0	0	0	0	0	0	3	1.0					x		x*	
23. Fountain Valley Reading Program		0	0	0	0	0	0	11	3.7								Testing in all skills except comprehension
24. Singer Visual Education		0	0	0	0	0	0	8	2.7					x		x	
25. Troll Cassettes & Filmstrips		0	0	0	0	0	0	7	2.3				x				Enjoyment
Other		31	17.7	3	1.1	19	5.6	7	2.3								
Teacher-made reading material		30	17.1	122	43.0	26	7.6	14	4.7								
Total instructional days sampled		175		284		342		301									

d. Grades 3 and 4: Math

Catch-Up Material	PIP Specified Core or Supplementary Material	Bloomington: Frequency of Use	Bloomington: Percent of Days Sampled	Brookport: Frequency of Use	Brookport: Percent of Days Sampled	Providence Forge: Frequency of Use	Providence Forge: Percent of Days Sampled	Wayne City: Frequency of Use	Wayne City: Percent of Days Sampled
Published math material									
1. Drill & Facts		2	1.1%	0	0%	0	0%	0	0
2. Holt Math Tape & Cassette		9	5.1	0	0	0	0	0	0
3. Holt Math Test		3	1.7	0	0	0	0	0	0
4. SRA Math Learning System†	Supp.	6	3.4	0	0	0	0	0	0
5. Singer Individualized Math	Supp.	14	8.0	24	8.5	17	5.0	14	4.7
6. Sullivan Math & Workbook	Core	11	6.3	74	26.1	40	11.7	14	4.7
7. Systems 80	Core	3	1.7	18	6.3	24	7.0	14	4.7
8. Tutor Computor	Core	5	2.9	42	14.8	0	0	45	15.0
9. Digitor		0	0	12	4.2	0	0	0	0
10. Math Drawer Worksheets		0	0	3	1.1	0	0	0	0
11. Houghton Mifflin Basic Facts & Skills		0	0	34	12.0	0	0	65	21.6
12. Houghton Mifflin Skill Sheets		0	0	33	11.6	0	0	38	12.6
13. Milton Bradley Fractions & Cassettes		0	0	10	3.5	0	0	0	0
14. Multiplication Records		0	0	7	2.5	0	0	0	0
15. Creative Filmstrips & Cassettes		0	0	0	0	5	1.5	0	0
16. Drill Pages		0	0	0	0	1	0.3	0	0
17. Drill Tapes		0	0	0	0	5	1.5	0	0
18. Mathiputer		0	0	0	0	3	0.9	0	0
19. Cyclo Teacher		0	0	0	0	0	0	5	1.7
20. SRA Arithmetic Fact Kit		0	0	0	0	0	0	12	4.0
21. Dominos (game)		0	0	0	0	0	0	6	2.0
22. Orbiting the Earth Game		0	0	0	0	0	0	6	2.0
Other		0	0	2	0.7	0	0	17	5.6
Teacher-made math material		41	23.4	26	9.2	13	3.8	34	11.3
Total instructional days sampled		175		284		342		301	

Note: Frequency tabulations were not made for Galax (for either reading or math) because of inadequate data.

*Sentence comprehension only.

†It was not clear in the hardware/software package whether the specified material referred to was the SRA Math Learning System or the SRA Mathematics Diagnosis.

e. Grades 5 and 6: Reading

Catch-Up Material	PIP Specified Core or Supplementary Material	Bloomington		Brookport		Providence Forge		Wayne City		Primer		Primer and Primary 1	Primary I	Primary I --Advanced	Primary II--Advanced		Comments
		Frequency of Use	Percent of Days Sampled	Frequency of Use	Percent of Days Sampled	Frequency of Use	Percent of Days Sampled	Frequency of Use	Percent of Days Sampled	Recognition of Sounds	Letter Recognition	Decoding	Structural Analysis	Vocabulary	Antonyms and Synonyms	Comprehension: Word, Sentence, and Paragraph	
Published reading material																	
1. Controlled reader		7	4.2%	0	0%	0	0%	0	0%					x		x	Also includes fluency
2. Creative Features Structural Analysis		4	2.4	0	0	28	8.2	0	0				x				
3. Open Court Correlated Language Arts & Reading Program	Supp.	11	6.5	0	0	18	5.3	0	0	x	x	x	x	x	x	x	
4. Sounds of Language Readers		7	4.2	0	0	0	0	0	0								Enjoyment
5. SRA Reading Laboratory:		9	5.4	0	0	12	3.5	0	0				x	x	x	x	
My Own Book		6	3.6	0	0	0	0	0	0	x				x		x	
6. SRA Reading Program	Supp.	13	7.3	0	0	0	0	0	0					x		x	
7. Vocabulary Development		6	3.6	0	0	0	0	0	0					x			
8. Systems 80	Core	4	2.4	5	1.9	15	4.4	8	2.7	x	x	x	x	x	x		
9. Random House Criterion Reading	Core	18	10.7	79	29.3	12	3.5	97	32.7								Testing in all skills
10. Adventure Trail		0	0	4	1.5	0	0	0	0								Enjoyment
11. Magic Seasons		0	0	6	2.2	0	0	0	0								Enjoyment
12. Minibike Film & Worksheet		0	0	10	3.7	0	0	0	0							x	Enjoyment
13. Phonics We Use (Old & New)		0	0	34	12.6	0	0	0	0	x	x	x	x	x			
14. Merrill Linguistic Readers		0	0	1	0.4	0	0	0	0	x	x	x	x	x	x	x	
15. Specific Skills Series	Supp.	0	0	17	6.3	69	20.2	8	2.7	x				x		x	
16. Mission Read		2	1.2	0	0	34	10.0	3	1.0					x		x	Enjoyment
17. Ginn Elementary English Series		0	0	0	0	11	3.2	0	0								English skills not listed (i.e., nouns & verbs)
18. Scholastic Individualized Reading	Core	3	1.8	0	0	11	3.2	0	0					x		x	
19. Words in Motion		0	0	0	0	37	10.9	0	0					x		x	
20. Core File (lessons correlated with Random House Criterion Reading)		0	0	0	0	0	0	16	5.4								Materials covering all skills
21. ESP Cassette Program		0	0	0	0	0	0	5	1.7			x					Varied skills, usually decoding
22. Filmstrips--Jim Handy		0	0	0	0	0	0	4	1.3					x		x	Enjoyment
23. Individual Cassette Learning Package		0	0	0	0	0	0	4	1.3								
24. Language Master Cards	Core	0	0	0	0	0	0	10	3.4	x	x	x		x	x		
25. Troll Cassettes & Filmstrips		0	0	0	0	0	0	6	2.0				x				Enjoyment
Other		10	6.0	1	0.4	16	4.7	12	4.0								
Teacher-made reading material		36	21.4	162	60.0	31	9.1	18	6.1								
Total instructional days sampled		168		270		341		297									

f. Grades 5 and 6: Math

Catch-Up Material	PIP Specified Core or Supplementary Material	Bloomington Frequency of Use	Bloomington Percent of Days Sampled	Brookport Frequency of Use	Brookport Percent of Days Sampled	Providence Forge Frequency of Use	Providence Forge Percent of Days Sampled	Wayne City Frequency of Use	Wayne City Percent of Days Sampled
Published math material									
1. Games		5	3.0%	0	0%	0	0%	0	0%
2. Holt Math Tapes and Cassettes		19	11.3	0	0	0	0	0	0
3. SRA Math Learning System*	Supp.	3	1.8	0	0	0	0	0	0
4. Singer Individualized Math	Supp.	13	7.7	16	5.9	44	12.9	8	2.7
5. Sullivan Math & Workbook	Core	3	1.8	20	7.4	17	5.0	31	10.4
6. Systems 80	Core	3	1.8	2	0.7	5	1.5	4	1.3
7. Tutor Computor	Core	4	2.4	18	6.7	0	0	30	10.1
8. BASE System		0	0	13	4.8	0	0	0	0
9. Milton Bradley Fractions & Cassettes		0	0	32	11.9	0	0	0	0
10. ICSS Fraction Kit		0	0	16	5.9	0	0	0	0
11. Math Drawer Worksheets		0	0	7	2.6	0	0	0	0
12. Houghton Mifflin Basic Facts & Skills		0	0	53	19.6	0	0	62	20.9
13. Houghton Mifflin Skill Sheets		0	0	8	3.0	0	0	6	2.0
14. Creative Filmstrips & Cassettes		0	0	0	0	8	2.3	0	0
15. Math Drills		0	0	0	0	1	0.3	0	0
16. Math Worksheets--Milliken		0	0	0	0	1	0.3	0	0
17. Sullivan Placement Test	Core	0	0	0	0	1	0.3	0	0
18. Educational Activities New Math Cassettes		0	0	0	0	0	0	4	1.3
19. Singer Visual Education		0	0	0	0	0	0	10	3.4
20. SRA Computapes		0	0	0	0	0	0	6	2.0
21. Triscore (game)		0	0	0	0	0	0	4	1.3
Other		0	0	5	1.9	0	0	30	10.1
Teacher-made math material		56	33.3	35	13.0	9	2.6	18	6.1
Total instructional days sampled		168		270		341		297	

Note: Frequency tabulations were not made for Galax because of inadequate data.

*The hardware/software package did not clearly indicate whether the material referred to was the SRA Math Learning System or the SRA Mathematics Diagnosis.

Table E-1 (Concluded)

g. Grades 7 and 8: Reading

Catch-Up Material	PIP Specified Core or Supplementary Material	Wayne City		Primer		Primer and Primary I	Primary I	Primary I--Advanced	Primary II--Advanced		Comments
		Frequency of Use	Percent of Days Sampled	Recognition of Sounds	Letter Recognition	Decoding	Structural Analysis	Vocabulary	Antonyms and Synonyms	Comprehension: Words, Sentence, and Paragraph	
Published reading material											
1. Cyclo Teacher		4	1.4%				x	x	x	x	
2. Probe (game)		3	1.0								
3. ESP Cassette Program		6	2.1			x					Varied skills, usually decoding
4. Fountain Valley Reading Program		7	2.4								Testing in all skills
5. Media Cassette		4	1.4								
6. Random House Criterion Reading	Core	96	33.3								Testing in all skills
7. Core File (lessons correlated with Random House Criterion Reading)		32	11.1								Materials covering all skills
8. Scholastic Skills Books		3	1.0	x	x	x	x	x			
9. Troll Cassettes & Filmstrips		5	1.7				x				Enjoyment
10. Tufabet Vocabulary Building Game		2	0.7	x	x			x			
Other		3	1.0								
Teacher-made reading material		26	9.0								
Total instructional days sampled		288									

h. Grades 7 and 8: Math

Published math material			
1. Houghton Mifflin Basic Facts & Skills		52	18.1%
2. Math Facts Division Game		2	0.7
3. Dominos (game)		4	1.4
4. Singer Individualized Math	Supp.	5	1.7
5. Singer Visual Education		6	2.1
6. Spinner Number Games		4	1.4
7. SRA Computapes		6	2.1
8. Sullivan Math	Core	28	9.7
9. Triscore (game)		5	1.7
10. Tutor Computor	Core	13	4.5
Other		15	5.2
Teacher-made math material		22	7.6
Total instructional days sampled		288	

Note: Frequency tabulations were not made for Bloomington, Brookport, or Providence Forge because these sites did not have students participating in the program at these grade levels; frequency tabulations were not made for Galax because of inadequate data.

Table E-2

MOST FREQUENTLY USED MATERIALS AND SKILLS EMPHASIZED IN CONQUEST PROJECTS, BY GRADE LEVEL

a. Grades 1 and 2: Reading

Conquest Material	FIP Specified Core or Supplementary Material	Benton Harbor		Cleveland		Primer		Primer and Primary I	Primary I	Primary I --Advanced	Primary II--Advanced		Comments
		Frequency of Use	Percent of Days Sampled	Frequency of Use	Percent of Days Sampled	Recognition of Sounds	Letter Recognition	Decoding	Structural Analysis	Vocabulary	Antonyms and Synonyms	Comprehension: Word, Sentence, and Paragraph	
Published reading material													
1. Games		28	10.0%	0	0 %								Varied skills*
2. D.C. Heath Workshops/Bookshops		54	19.3	0	0		x	x		x		x	
3. Merrill Reading Skill Text	Core	100	35.7	0	0	x	x	x	x	x			
4. Phonovisual Phonics		79	28.2	0	0	x	x	x	x				
5. Patterns, Sounds & Meaning		48	17.1	0	0	x	x	x					
6. Read story/book		52	18.6	0	0								Enjoyment
7. Phonovisual Consonant Workbook		32	11.4	0	0	x	x	x					
8. McGraw-Hill Programmed Reading	Core	201	71.8	36	15.5	x	x	x	x	x			
9. Systems 80	Core	58	20.7	64	27.5	x	x	x	x	x			
10. Dolch Vocabulary Words (WCRC)†	Core	69	24.6	82	35.2					x			
11. Bowmar Primary Reading Series	Core	0	0	70	30.0								Enjoyment
12. Reader's Digest Individual Skill Builders	Supp.	0	0	54	23.2							x	
13. Specific Skills Series	Core	0	0	44	18.9	x				x		x	
14. Ideal Tapes & Worksheets		0	0	31	13.3			x					Varied skills, usually decoding
15. Phonics We Use		0	0	39	16.7	x	x	x	x				
16. Read, Study, Think	Core	0	0	19	8.2	x						x	
17. Steck-Vaughn Individualized Directions in Reading		0	0	16	6.9			x	x			x	
Other		138	49.3	95	40.8								
Teacher-made reading material		16	5.7	155	66.5								
Total instructional days sampled		280		233									

b. Grades 3 and 4: Reading

Conquest Material	FIP Specified Core or Supplementary Material	Benton Harbor		Cleveland		Primer		Primer and Primary I	Primary I	Primary I --Advanced	Primary II--Advanced		Comments
		Frequency of Use	Percent of Days Sampled	Frequency of Use	Percent of Days Sampled	Recognition of Sounds	Letter Recognition	Decoding	Structural Analysis	Vocabulary	Antonyms and Synonyms	Comprehension: Word, Sentence, and Paragraph	
Published reading material													
1. D.C. Heath Workshops/Bookshops		50	20.8%	0	0 %		x	x		x		x	
2. Merrill Reading Skill Text	Core	61	25.4	0	0	x	x	x	x	x			
3. Patterns, Sounds & Meaning		44	18.3	0	0	x	x	x	x	x			
4. Phonovisual Phonics		49	20.4	0	0	x	x	x	x	x			Encoding as part of decoding skills
5. Language Master Cards		25	10.4	0	0			x		x			Varied skills, usually decoding & vocabulary
6. SRA Reading Laboratory	Core	30	12.5	0	0				x	x	x	x	
7. Specific Skills Series	Core	56	23.3	89	33.7	x				x	x	x	
8. McGraw-Hill Programmed Reading	Core	181	75.4	48	18.2	x	x	x		x			
9. Systems 80	Core	45	18.8	22	8.3	x	x	x	x	x			
10. Dr. Spello (WCRC)†	Core	34	14.1	4	1.5	x	x	x	x	x			
11. Dolch Vocabulary Words (WCRC)†	Core	0	0	24	9.1					x			
12. Read, Study, Think	Core	8	3.3	16	6.1	x						x	
13. Scholastic Skills Books		0	0	43	16.3	x	x	x	x				
14. Phonics We Use		0	0	54	20 5	x	x	x	x				
15. Controlled Reader	Core	0	0	21	7.8					x		x	Also includes fluency & speed
16. Merrill Phonics Skill Text	Core	0	0	18	6.8	x	x	x	x	x			
17. Reader's Digest Individual Skill Builders	Supp.	0	0	83	31.4					x	x	x	
Other		155	64.6	104	39.4								
Teacher-made reading material		48	20.0	206	78.0								
Total instructional days sampled		240		264									

Note: Frequency tabulations could not be made for Gloversville because of inadequate data.

*It was unclear what skills were covered due to lack of knowledge about games used.

†WCRC = Webster Classroom Reading Clinic, which includes: Conquests in Reading, Dr. Spello, Dolch Basic Sight Vocabulary, and others. The components were dealt with individually because of the different skills that each covered.

Table E-2 (Concluded)

c. Grades 5 and 6: Reading

Conquest Material	PIP Specified Core or Supplementary Material	Benton Harbor		Cleveland		Primer		Primer and Primary I	Primary I	Primary I--Advanced	Primary II--Advanced		Comments
		Frequency of Use	Percent of Days Sampled	Frequency of Use	Percent of Days Sampled	Recognition of Sounds	Letter Recognition	Decoding	Structural Analysis	Vocabulary	Antonyms and Synonyms	Comprehension: Word, Sentence, and Paragraph	
Published reading material													
1. Games		20	8.8%	0	0 %								Varied skills*
2. D.C. Heath Workshops/Bookshops		57	25.0	0	0		x	x		x		x	
3. Merrill Reading Skill Text	Core	30	13.2	0	0	x	x	x	x	x			
4. Patterns, Sounds & Meaning		99	43.4	0	0	x	x	x	x	x			
5. New Spelling Goals		21	9.2	0	0	x	x	x	x	x			
6. SRA Reading Laboratory	Core	39	17.1	0	0				x	x	x	x	
7. Specific Skills Series	Core	85	37.3	93	33.8	x				x		x	
8. McGraw-Hill Programmed Reading	Core	164	71.9	90	32.7	x	x	x	x	x			
9. Systems 80	Core	31	13.6	73	26.5	x	x	x	x	x			
10. Merrill Phonics Skill Text	Core	4	1.8	90	32.7	x	x	x	x	x			
11. Reader's Digest Individual Skill Builders	Supp.	4	1.8	39	14.2					x	x	x	
12. Dr. Spello (WCRC)†	Core	64	28.1	28	10.2	x	x	x	x	x			Encoding as part of decoding skills
13. Conquests in Reading (WCRC)†	Core	19	8.3	54	19.6	x	x	x	x	x			
14. Steck-Vaughn Individualized Directions in Reading		0	0	47	17.1			x	x			x	
15. Phonics We Use		0	0	35	12.7	x	x	x	x		x		
16. Tach-X		0	0	24	8.7		x			x			Also includes fluency & speed
Other		162	71.0	150	54.5								
Teacher-made reading material		[illegible]	[illegible]	21	7.6								
Total instructional days sampled		228		275									

Note: Frequency tabulations could not be made for Gloversville because of inadequate data.

*It was unclear what skills were covered due to lack of knowledge about games used.

†WCRC = Webster Classroom Reading Clinic, which includes: Conquests in Reading, Dr. Spello, Dolch Basic Sight Vocabulary, and others. The components were dealt with individually because of the different skills that each covered.

Table E-3

MOST FREQUENTLY USED MATERIALS AND SKILLS EMPHASIZED IN HIT PROJECTS, BY GRADE LEVEL

a. Grade 6: Reading

HIT Material	PIP Specified Core or Supplementary Material	Lexington		Primer		Primer and Primary I	Primary I	Primary I --Advanced	Primary II--Advanced		Comments
		Frequency of Use	Percent of Days Sampled	Recognition of Sounds	Letter Recognition	Decoding	Structural Analysis	Vocabulary	Antonyms and Synonyms	Comprehension: Word, Sentence, and Paragraph	
Published reading material											
1. Adventuring in the City		2	1.1%								Enjoyment
2. Conquests in Reading*		34	18.8	x	x	x	x	x	x		
3. Games		10	5.5								Varied skills†
4. Phonics We Use		11	6.1	x	x	x	x		x		
5. New Phonics We Use		3	1.7	x	x	x	x		x		
6. Read story/book (unspecified)		2	1.1								Enjoyment
7. Remedial Reading Drills	Core	62	34.3			x					
8. Stories of the Inner City	Core	20	11.0								Enjoyment
Other‡		4	2.2								
Teacher-made reading material		16	8.8								
Total instructional days sampled		181									

Note: Frequency tabulations were not made for Olean because it did not have students participating in the program at this grade level.

*This material was included in the hardware/software packet by mistake, but Lexington continued to use the material during the second year of the program because they had problems obtaining the Sullivan Programmed Reading materials.

†It was unclear what skills were covered due to lack of knowledge about games used.

‡Includes Pay Day activities; Pay Day involves use of a bank book, which is not a published material but is specified in the PIP.

Table E-3 (Concluded)

b. Grades 7-9: Reading

HIT Material	PIP Specified Core or Supplementary Material	Olean: Frequency of Use	Olean: Percent of Days Sampled	Lexington: Frequency of Use	Lexington: Percent of Days Sampled	Primer: Recognition of Sounds	Primer: Letter Recognition	Primer and Primary I: Decoding	Primary I: Structural Analysis	Primary I--Advanced: Vocabulary	Primary II--Advanced: Antonyms and Synomyms	Primary II--Advanced: Comprehension: Word, Sentence, and Paragraph	Comments
Published reading material													
1. Adventuring in the City		0	0 %	9	4.3%								Enjoyment
2. Conquests in Reading		0	0	36	17.1	x	x	x	x		x		
3. Phonics We Use		0	0	19	9.0	x	x	x	x		x		
4. New Phonics We Use		0	0	5	2.4	x	x	x	x		x		
5. Stories of the Inner City	Core	0	0	21	10.0								
6. Games		4	2.6	3	1.4								Enjoyment Varied skills*
7. Remedial Reading Drills	Core	38	24.7	78	37.1			x					
8. SRA Reading Laboratory		5	3.2	0	0				x	x	x	x	
9. Sullivan Comprehension Readers		64	41.6	0	0					x		x	
10. Sullivan Programmed Reading	Core	97	63.0	0	0	x	x	x	x	x			
Other		8	5.2	8†	3.8								
Teacher-made reading material		0	0	8	3.8								
Total instructional days sampled		154		210									

c. Grades 7-9: Math

Published math material		Olean‡	
1. Self-Teaching Flashcards in Addition and Subtraction	Core	20	10.4%
2. Self-Teaching Flashcards in Division	Core	46	24.0
3. Self-Teaching Flashcards in Multiplication	Core	56	29.2
4. SRA Arithmetic Fact Kit		2	1.0
5. Sullivan Math§	Core	160	83.3
6. Pay Day		3	1.6
Total instructional days sampled		192	

*It was unclear what skills were covered due to lack of knowledge about games used.

†Includes five Pay Day activities; Pay Day involves use of a bank book, which is not a published material but is specified in the PIP.

‡Frequency tabulations were not made for Lexington because of inadequate data relative to math instruction.

§Although the Sullivan Math Program is not specified in the hardware/software packet, the Sullivan Placement Test and Test Booklet are specified and are part of the Sullivan Math Program.

Table E-4

MOST FREQUENTLY USED MATERIALS AND SKILLS EMPHASIZED IN IRIT PROJECTS, BY GRADE LEVEL

a. Grades 3 and 4: Decoding Instruction

PIP-Specified Core or Supplementary Material	Bloomington		Oklahoma City		Schenectady		Primer		Primer and Primary I	Primary I	Primary I--Advanced	Primary II--Advanced		Comments
	Frequency of Use	Percent of Days Sampled	Frequency of Use	Percent of Days Sampled	Frequency of Use	Percent of Days Sampled	Recognition of Sounds	Letter Recognition	Decoding	Structural Analysis	Vocabulary	Antonyms and Synonyms	Comprehension: Word, Sentence, and Paragraph	
	12	6.9%	0	0 %	0	0 %								Enjoyment, listening skills
	27	15.6	0	0	0	0	x	x	x	x	x	x	x	
	38	22.0	0	0	0	0								
Core	96	55.5	0	0	0	0	x	x	x	x	x	x		
	7	4.0	0	0	0	0								
	29	16.8	0	0	0	0								Enjoyment & fluency
	12	6.9	0	0	0	0					x			
Supp.	15	8.7	14	6.2	1	1.2			x		x			Varied skills, usually vocabulary & decoding
Core	57	32.9	63	28.0	0	0	x	x	x	x		x		
	34	19.7	17	7.6	8	9.4	x	x	x	x		x		
	0	0	19	8.4	0	0					x			
	0	0	43	19.1	0	0				x	x		x	
	0	0	161	71.6	0	0					x			
	0	0	50	22.2	0	0	x	x	x					
	0	0	32	14.2	0	0					x	x	x	
Core	0	0	58	25.8	0	0	x				x	x	x	
	0	0	0	0	5	5.9								Testing in all skills
	0	0	0	0	5	5.9								Testing in all skills
	0	0	0	0	5	5.9	x	x	x					
	0	0	0	0	5	5.9	x	x	x	x	x			Enjoyment
	0	0	0	0	11	12.9	x	x	x	x				
	0	0	0	0	15	17.6			x					
	0	0	0	0	10	11.8			x					
	0	0	0	0	14	16.5			x					
	0	0	0	0	8	9.4			x					
	21	12.1	20	8.9	20	23.5								
	118	68.2	7	3.1	0	0								
	173		225		85									

Table E-4 (Continued)

b. G. [illegible] and 4: Individualized Reading Instruction

IRIT Reading Material (individualized reading instruction)	PIP-Specified Core or Supplementary Material	Bloomington: Frequency of Use	Bloomington: Percent of Days Sampled	Oklahoma City: Frequency of Use	Oklahoma City: Percent of Days Sampled	Schenectady: Frequency of Use	Schenectady: Percent of Days Sampled	Primer: Recognition of Sounds	Primer: Letter Recognition	Primer and Primary I: Decoding	Primary I: Structural Analysis	Primary I--Advanced: Vocabulary	Primary II--Advanced: Antonyms and Synonyms	Primary II--Advanced: Comprehension: Word, Sentence, and Paragraph	Comments
Published material															
1. Bear's Picnic		11	4.4%	0	0 %	0	0 %								Enjoyment
2. Frances Series		7	2.8	0	0	0	0								Enjoyment
3. Walt Disney Story Records	Core	13	5.2	0	0	0	0								Enjoyment, listening skills
4. Individualized Cassette Learning Package		25	10.0	0	0	0	0								Enjoyment, listening skills
5. Random House Reading Program	Core	127	50.8	0	0	0	0				x	x	x	x	
6. Scholastic Individualized Reading	Core	90	36.0	5	2.2	0	0				x	x	x	x	
7. Specific Skills Series	Core	0	0	30	13.3	0	0	x	x	x	x	x	x	x	
8. Tape Recorder with Books		0	0	77	34.2	0	0								Enjoyment, listening skills
9. Trade Books		0	0	118	52.4	0	0								Enjoyment
10. Work grading with Specific Skills Series & Random House Reading Materials*		0	0	36	16.0	0	0				x	x	x	x	
11. Auto Vance Films		0	0	39	17.3	0	0								Enjoyment, listening skills
12. Continuous Progress Laboratory for Language Arts		0	0	75	33.3	0	0				x	x	x	x	
13. Holtman Language Arts Reading Program		0	0	156	69.3	0	0					x	x	x	
14. Language Master Card Program	Supp.	0	0	13	5.8	0	0			x		x			Varied skills, usually decoding & vocabulary
15. Newspaper Stories		0	0	24	10.7	0	0							x	Writing skills
16. Adventures in Glen Series		0	0	0	0	7	8.5								Enjoyment
17. Book Bags		0	0	0	0	19	23.2								Enjoyment, listening skills
18. Dr. Seuss Series		0	0	0	0	9	11.0								Enjoyment
19. Mr. & Mrs. Bumbo Series		0	0	0	0	12	14.6								Enjoyment
20. Recreational Reading		0	0	0	0	9	11.0								Enjoyment
21. Taylor Filmstrips: Tell-Me-A Story Library		0	0	0	0	6	7.3							x†	Sequencing, listening & speaking skills
Other		138	55.2	0	0	27	32.9								
Teacher-made reading material		52	20.8	8	3.6	1	1.2								
Total instructional days sampled		250		225		82									

*Work grading is not a published material but is a specific and separate use of the published materials listed.

†Word and sentence comprehension only.

Table E-4 (Concluded)

c. Grades 3 and 4: Vocabulary and Comprehension

IRIT Reading Material (vocabulary and comprehension)	PIP-Specified Core or Supplementary Material	Bloomington		Oklahoma City		Primer		Primer and Primary I	Primary I	Primary I--Advanced	Primary II--Advanced		Comments
		Frequency of Use	Percent of Days Sampled	Frequency of Use	Percent of Days Sampled	Recognition of Sounds	Letter Recognition	Decoding	Structural Analysis	Vocabulary	Antonyms and Synonyms	Comprehension: Word, Sentence, and Paragraph	
Published material													
1. Bowmar Highway Holiday Series		48	26.2%	0	0 %								Enjoyment
2. Continuous Progress in Spelling		74	40.4	0	0		x	x	x	x	x		
3. Random House Criterion Reading		35	19.1	0	0								Testings in all skills
4. Phoenix Reading Series		6	3.3	0	0			x		x		x	
5. Poems		4	2.2	0	0								Enjoyment & word patterns
6. Story Picture		11	6.0	0	0								Enjoyment
7. Controlled Reader	Supp.	11	6.0	20	8.9					x		x	
8. Reader's Digest Individual Skill Builders	Core	10	5.5	56	24.9					x		x	
9. Specific Skills Series	Core	26	14.2	58	25.8					x		x	
10. SRA Reading Laboratory	Core	25	13.7	10	4.4				x	x	x	x	
11. Houghton Mifflin Readers & Workbook		0	0	48	21.3	x	x	x	x	x	x	x	
12. Language Center 2		0	0.	26	11.6								
13. Macmillan Basals & Workbook		0	0	43	19.1	x	x	x	x	x	x	x	
14. Merrill Linguistic Readers	Core	0	0	81	36.0	x	x	x	x	x	x	x	
15. Taylor Filmstrips: Tell-Me-A Story Library		0	0	19	8.4							x*	Sequencing, listening skills & speaking skills
16. Typing—McGraw-Hill		0	0	77	34.2								Language arts skills
Other		19	10.4	44	19.6								
Teacher-made reading material		17	9.3	34	15.1								
Total instructional days sampled		183		225									

Note: Frequency tabulations could not be made for Schenectady because no data were available.

* Word and sentence comprehension only.

Appendix F

MAT SUBTEST ITEMS USED IN REGRESSION ANALYSES

Appendix F

MAT SUBTEST ITEMS USED IN REGRESSION ANALYSES

Mat subtest items displayed in Table E-1 are comparable between MAT batteries shown. At the fourth and eighth grades, the items compared do not represent the total possible comparable items, but [illegible] the comparable items among those selected as relevant to the [illegible] curriculum only.

Table F-1

PARALLEL MAT SUBTEST ITEMS RELEVANT TO PIP CURRICULUM

All Subtests: Grade 1 Only

(a)	Primer: Listening for Sounds	Primary I: Word Analysis
	34	4
	36	2
	37	21
	39	5
(b)	Reading	Word Knowledge
	16	3
	17	5
(c)	Reading	Reading
	None	[illegible]
(d)	Numbers	Mathematics
	1	3
	3	2
	4	7
	5	10
	6	12
	8	4
	10	32
	12	33
	13	9
	14	11
	17	17
	19	16
	21	39
	22	38
	23	36
	24	40
	25	41
	26	48
	27	49
	28	50
	30	37
	31	43
	32	44
	34	47

Word Knowledge Subtest: All Grades Except Grade 1

(a)	Primary I	Primary II
	23	14
	32	11
(b)	Primary II	Elementary
	None	None
(c)	Elementary	Elementary
	1-50	1-50
(d)	Elementary	Intermediate
	None	None
(e)	Intermediate	Advanced
	None	None
(f)	HIT Advanced	HIT Advanced
	1-50	1-50
(g)	R-3 Advanced	R-3 Advanced
	1-50	1-50

Table F-1 (Continued)

Reading Subtest: All Grades Except Grade 1

(a)	Primary I	Primary II
	29	15
	30	14
	31	16
	32	18
	38	25
	39	30
	40	26
	42	29

(b)	Primary II	Elementary
	22	14
	23	11
	24	12
	25	1
	26	3
	29	4
	36	5
	37	6
	38	7
	39	8
	40	9
	42	18
	44	17

(c)	Elementary	Elementary
	1-28	1-28

(d)	Elementary	Intermediate
	29	5
	30	10
	31	8
	32	7
	34	9
	36	3
	38	1
	39	4
	40	24
	42	22
	43	28
	45	26

(e)	Intermediate	Advanced
	16	20
	18	22
	19	21
	36	13
	39	14
	40	11
	41	5
	42	7
	43	6
	44	8
	45	9

(f)	HIT Advanced	HIT Advanced
	1-15	1-15
	30-37	30-37

(g)	R-3 Advanced	R-3 Advanced
	1-15	1-15
	30-37	30-37

Table F-1 (Continued)

Mathematics Computation Subtest: All Grades Except Grade 1

(a) Primary I: Mathematics, Part B	Primary II
36	13
38	1
39	2
41	7
42	3
46	12
48	9
50	8
53	5
55	16
56	29
57	6
58	14
59	[illegible]
61	[illegible]

(b) Primary II	Elementary
3	2
6	10
9	4
10	9
11	8
13	1
15	11
19	3
22	16
23	15
26	12
28	17
29	6
30	19
31	34
33	14

(c) Elementary	Elementary
1-50	1-40

(d) Elementary	Intermediate
2	15
11	2
12	6
14	22
17	4
22	5
24	3
27	19
28	7
29	17
30	13
31	1
34	34
36	18
40	11

(e) Intermediate	Advanced
5	1
10	5
11	6
12	14
13	4
15	2
19	9
20	7
23	12
26	3
27	33
28	10
29	13
30	18
31	19
33	22
36	16
38	37
40	28

(f) HIT Advanced	HIT Advanced
1-12	1-12
14-16	14-16
18-20	18-20
22-24	22-24
26-31	26-31
33	33
35-37	35-37

(g) R-3 Advanced	R-3 Advanced
1-25	1-25
27-40	27-40

Table F-1 (Continued)

Mathematics Concepts Subtest: All Grades Except Grade 1		
(a)	Primary I Mathematics, Part A	Primary II
	1	6
	2	8
	3	1
	4	17
	6	3
	9	13
	15	11
	16	12
	18	15
	19	29
	21	29
	23	22
	25	24
	26	18
	28	39
	35	35
(b)	Primary II	Elementary
	3	4
	6	2
	7	8
	8	1
	10	5
	11	10
	13	21
	14	27
	17	3
	21	13
	22	20
	25	14
	26	18
	29	29
	30	17
	31	33
	36	25
	38	16
(c)	Elementary	Elementary
	1	1
	2	2
	3	3
	7	7
	8	8
	20	20
	21	21
	23	23
	26	26
	35	35
(d)	Elementary	Intermediate
	12	15
	13	2
	15	3
	18	5
	19	1
	23	4
	25	19
	27	10
	28	20
	29	9
	31	7
	32	21
	37	29
	38	31
	39	6
(e)	Intermediate	Advanced
	2	6
	3	10
	12	4
	13	1
	17	3
	18	2
	19	18
	23	26
	26	14
	30	27
	34	12
	36	25
	37	20
	39	16
(f)	HIT Advanced	HIT Advanced
	1	1
	4	4
	20	20
(g)	R-3 Advanced	R-3 Advanced
	1-9	1-9
	11-13	11-13
	15-40	15-40

Table F-1 (Concluded)

Mathematics Problem Solving Subtest: All Grades Except Grade 1

(a)

Primary I: Mathematics, Part A	Primary II
17	1
22	3

(b)

Primary II	Elementary
18	2
19	1
21	4
22	5
23	3
26	15
28	8
30	11
31	28
32	12
33	31
34	32

(c)

Elementary	Elementary
1	1
2	2
3	3
4	4
6	6
10	10
17	17
25	25

(d)

Elementary	Intermediate
5	1
8	2
13	4
16	3
17	12
20	22
23	6
24	9
26	15
28	8
30	7

(e)

Intermediate	Advanced
6	1
9	4
11	3
15	5
10	2
12	8
16	21
17	7
19	10
21	17
28	9
29	6
30	27

(f)

HIT Advanced	HIT Advanced
1	1
2	2
6	6
12	12

(g)

R-3 Advanced	R-3 Advanced
1-2	1-2
5-8	5-8
10	10
12-17	12-17
19	19
21-29	21-29
31-35	31-35

Appendix G

ANALYSIS OF CORRESPONDENCE BETWEEN THE MAT
AND FOURTH AND EIGHTH GRADE CURRICULA

Appendix G

ANALYSIS OF CORRESPONDENCE BETWEEN THE MAT AND FOURTH AND EIGHTH GRADE CURRICULA

Our analyses showed that at a gross level the PIP curricula covered skills tested by the MAT. However, these analyses are not entirely satisfactory because they were based on global judgments, without regard to the placement of students relative to the curriculum materials. In this section we report analyses that are more conservative than those reported above, in the sense that we are much more careful about specifying what we mean by a skill being "tested by the MAT."

To determine if the MAT is truly testing the curriculum, we must be sure that the skills needed to answer an item correctly are "taught" in the curriculum materials. We must also be sure that we have criteria for deciding that students covering those curriculum materials would have learned those skills.

Information was available in the SOIs at a fairly fine level of detail on what lessons students had covered. We therefore had a reasonably detailed picture of what materials were used in teaching, but could not say exactly what was taught because materials could be used to teach several points. We hoped that one of the points taught was the one that the author of the materials intended; if so, we could much more effectively assess the relevance of the MAT to PIP outcomes, and could restrict our statistical analyses to just those items that were relevant to the curriculum.

Our objective was to establish a correspondence between the PIP-specified-and-used curriculum and the MAT, based on our idea of what skills were necessary to learn an item (i.e., to answer it correctly). Mapping this correspondence entailed two tasks: (1) analysis of the skills required for a correct response to each test item, and development of the rules or criteria for deciding what in the curriculum would be an exemplar of that set of skills and (2) development of procedures for searching the curriculum materials to find units that satisfied the rules. In this section, then, the methods for determining congruence between what was tested and what was taught are presented.

Our use of these methods was limited to the Elementary and the Advanced levels of the MAT and to those PIP projects with students in the fourth and the eighth grades (except for R-3). Only at these grades were the children tested in both the fall and spring on the same form and level of the MAT. We did not distinguish between the specified-and-used curriculum of fourth grade Catch-Up, Conquest, and IRIT because they were about the same. The eighth grade curriculum is the same as that for HIT.

Skills Analysis of the Elementary and Advanced MAT

Analysis of the MAT began with an examination of each item. Our strategy for determining what skills would be necessary for answering that item correctly was to ask ourselves what would "teach" that particular item as presented in that particular format. Generally we found that each test item required a combination of skills including responding appropriately to the item format. We tried to imagine the kind of curriculum unit that would give a student experience with the set of skills and knowledge he would need to ensure a correct response.

In formulating the rules for determining whether the test item had been "taught," we were extremely literal about the features of the item. We included in our rules, or criteria for declaring a match between curriculum and test item, all the features we felt were essential for the student to answer the item correctly. We wished to take this conservative approach so that, when eventually we included a test item in our analysis of project effectiveness, we could assert that students covered the materials appropriate to passing that item.

Word Knowledge Subtest

The rules for the Word Knowledge subtest were quite unambiguous. Generally, for each test item, the two words in the stem of the item and the correct answer word had to be found in the curriculum for a match to be declared. (For example, Item 3 in the Elementary subtest is "happy means glad." In the stem of the item, "happy" is the target word, and "means" is the context word. The answer word is "glad.") Modified versions of any of the three words could not change the meaning of the word. We used the principle of "near transfer" as a guide to limit which modifications would be acceptable.* For example, "dependable" and "depend" would be considered acceptable because the idea of being reliant on another is basic to both words. "Please" versus "pleasing" would not be credited because "please" usually functions as a way to express politeness, whereas "pleasing" connotes giving pleasure.

---

*This principle assumes hardly any generalization of the skills.

We felt that all three words in an item were important because to get to the answer or critical word from the target word the student must understand the context word. For example, to know that "night" is the opposite of "day," the student must know the meaning of the context word "opposite." We therefore concluded that the student must have had experience with all three words.

For the Word Knowledge subtest we claimed that a curriculum unit taught a word only if the particular word and its meaning were treated in a well-marked exercise and if several practice items were included in the exercise. It was not practical to set a threshold, such as two or four practice items within an exercise, because of the diversity of presentations in the curriculum materials. For word knowledge items, it was also necessary that the meaning of the word be singled out for attention in the curriculum unit or exercise and that the student be required to determine the meaning of the word from some contextual clues.

Reading Subtest

In the Reading subtest, a pupil is required to read a passage and respond to several questions about the passage. Exhibit G-1 is an example from the Elementary MAT. Rules for MAT items were based on two kinds of features--one describing the passage and the other specifying the type of question posed about the passage. Although we believe knowledge of the content of a passage would sometimes permit the item to be answered correctly even when the passage was not read or comprehended (Tuinman, 1973-74), we could not conceive of any way to develop rules for matching content in the test passages with content in the curriculum. Because of our conservative strategy we were not particularly concerned with whether students could pass items without having covered relevant curriculum material. Rather we wished to claim that certain curriculum materials contained all the requisite skills for certain items and thus that students who could not formerly pass the item should now be able to pass.

Passages used to test reading comprehension vary on several dimensions besides content. They are generally made more difficult to comprehend by (1) containing more words that either occur infrequently in the students' experience or are abstract or complex in meaning, (2) employing more phrases or clauses requiring ideas to be temporarily stored in memory before the message is complete, or (3) lengthening the passage so that attentional skills, memory, or search skills are taxed. The MAT publishers stated that variation in the reading passages occurred along three dimensions: vocabulary level, syntactic complexity, and length of

EXHIBIT G-1. SAMPLE PASSAGE AND ITEMS FROM THE ELEMENTARY MAT

Each year on November 5, people in England celebrate a special holiday. The holiday, Guy Fawkes Day, is enjoyed by both children and older people. Huge bonfires are lit, and in the evening, children set off fireworks. Fawkes lived more than 350 years ago. He took part in the famous "Gunpowder Plot" against the ★government. The English **still** celebrate the day because the plot was discovered before anyone was hurt.

17 The best name for this story would be —

- Ⓐ Holidays
- Ⓑ An English Holiday
- Ⓒ The Life of Guy Fawkes
- Ⓓ Bonfires

18 Children set off fireworks —

- Ⓔ all day
- Ⓕ on bonfires
- Ⓖ after November
- Ⓗ in the evening

19 Instead of **still**, in the last sentence, you could say —

- Ⓐ calm
- Ⓑ quietly
- Ⓒ continue to
- Ⓓ at rest

20 The gunpowder plot probably took place —

- Ⓔ in British legend
- Ⓕ on a holiday
- Ⓖ in the spring
- Ⓗ in the early 1600's

passage. We quantified these dimensions to create rules for matching MAT test passages with PIP curriculum passages.

The Spache and Dale-Chall readability formulas (Spache; Dale and Chall, 1948) were used to measure the first dimension (i.e., vocabulary level). These formulas generate a quantitative measure in the form of a grade level for which vocabulary is appropriate. The indices are primarily dependent on vocabulary, but include some adjustment for average sentence length in the passage. The number of words in a passage that are not included on a master list are counted. The greater the number of exclusions, the more difficult (higher grade level) the passage is rated. The Spache index was used for passages up to 4th grade level, and the Dale-Chall index was used for 5th to 11th grade material. The Flesch formula (see Klare, 1974-75) has been used for fourth-fifth grade, but its reliability has been questioned. Therefore, a weighted combination of the Spache and Dale-Chall formulas was used to quantify readability at approximately the fourth to fifth grade level.*

These two formulas, among the numerous readability indexing techniques, predict reading level most reliably (Klare, 1974-75). The test publishers used a simple noun count, but this technique underestimates difficulty (Klare). A computer program (Judd, 1975) was used for this measure to reduce tedious hand calculations.

For syntactic complexity--the second dimension of passages to be indexed--a sensitive measure was considered. This measure, the unit of which is called a T-unit (an independent and linked dependent clause), has accurately discriminated among children's writing as well as reading passages in norm-referenced tests (Hunt, 1965; Calfee, 1975). Another cluster of syntactic dimensions, including occurrence of certain syntactic features in words (Golub and Kidder. 1974), was also examined. Both techniques were rejected because the amount of reading materials in the PIP curricula was too extensive to analyze by these methods within the time frame of the project.

---

*This weighting was constructed by using the length of the passage (say 65 words) as a percentage to adjust the Dale-Chall (say 9.5 grade level). The product (grade level) was averaged with the Spache grade level. The caveat appropriate for never averaging grade level equivalent test scores does not apply to these grade level scores. The grade level scores in readability formulas are not projections of performance; they characterize materials. However, like grade level equivalency scores, they are viewed as approximations, indices of a vocabulary level of reading comprehension.

A simple measure of sentence length (number of words in passage divided by number of sentences in passage) was chosen to reflect syntactic complexity. The longer the sentence, it was assumed, the more cognitive processing is required to understand it. This assumption, while weaker than we would have liked, holds up frequently enough to justify its use.

Length of the passages that the student must read was the third dimension, or feature, measured. In most tests, length of a passage discriminates among students' test scores because the test time is limited and students who read faster cover more items. Because the student rarely encounters severe time limits in the classroom, the test is measuring performance for which he has not been trained. In addition, his skills cannot be tapped if he is unable to finish the long passages due to time constraints. Consequently, the length of each passage measured in number of words was viewed as a critical feature.

Test passages had different combinations of the three measures. Sometimes, a long passage had a lower vocabulary level than a shorter passage, or a short passage contained long sentences. Our purpose in analyzing the test required that the three quantitative scores for each MAT test passage be compared with the corresponding scores for each sample of the PIP curriculum. Table G-1 shows the ranges on the three measures for matching PIP curriculum passages with the Elementary and Advanced levels of the MAT. The passages were matched when a PIP curriculum unit had scores that fell within the same ranges as the MAT test passage on all three measures (or within the same range on two measures and higher on one). Unlike the other skill areas, reading comprehension was considered cumulative. "Taught" here was defined as "pass beyond." If a curriculum unit contained passages of greater difficulty, it was concluded that it "taught" passages of lesser difficulty. For example, if a student had read a passage with a vocabulary measure of 3.5, a sentence measure of 10.3, and a passage length measure of 200, we assumed he would be able to read passages whose measures were lower than these, whether or not we could demonstrate that he had actually read passages at that level. Thus, if we found that a reading curriculum of sixth grade level was used throughout the year, third grade items were considered covered as well.

Four types of questions were associated with the test passages on the MAT Reading subtest. These were "main idea," "literal," "inferential," and "word-in-context" questions. Again applying the concept of near transfer, we decided the curriculum fragment must include both a reading passage at the appropriate level and one or more of the four types of questions. For example, if a curriculum passage was of an appropriate level for the Guy Fawkes reading item we displayed earlier as

Table G-1

RANGES ON THREE INDICES FOR MATCHING READING PASSAGES IN PIP CURRICULUM WITH READING PASSAGES IN THE MAT

Passage Feature	Test Level	
	Elementary	Advanced
Vocabulary range*	-2.5 2.6-3.2 3.3-3.9 4.0-4.2 4.3-4.6 4.7-5.6 5.7-5.8 5.9-6.4 6.5+	-7.6 7.7-8.1 8.2-8.8 8.9-9.1 9.2-10.7 10.8-11.5 11.5+
Mean sentence length (number of words in passage divided by number of sentences in passage)	8.7-9.2 9.8-10.6 10.7-10.8 10.9-11.0 11.1-11.2 11.3-13.1 13.2-13.6 13.7-14.8 14.9+	12.2-12.6 12.7-14.4 14.5-14.8 14.9-15.0 15.1-16.3 16.4-17.3 17.4-18.5 18.6-18.8 18.9+
Passage length (in number of words)	55-65 66-77 78-79 80-82 83-89	105-149 150-159 160-202 203-261 262-303 304-313 314+

*Elementary level, Spache/Dale-Chall; advanced level, Dale-Chall.

Exhibit G-1, and if that curriculum passage had only inferential questions associated with it, we counted it a match only for Item 20 in Exhibit G-1. If a main idea question was also associated with the passage, we also counted a match for Item 17.

The other requirement for matching the reading comprehension items with a curriculum material was that the material be one that required a student to read the curriculum passages and the questions to himself. We felt that if a student used a tape with a book or conferenced with the teacher, he would not have been given the skills needed to pass the test.

Mathematics Computation Subtest

Like the subtest for word knowledge, the Mathematics Computation subtest is quite straightforward. We reasoned that the math computation items would probably appear vertically in almost the same manner in the curriculum; the only variation would be in the actual value of the numbers. Our rules which were stated as questions that had to be answered for both the curriculum item and the test item, specified the kinds of skil required for each item. Among the ten math computation questions were the following: (1) What operation is being performed? (2) What types of numbers are being used? (3) Does the operation involve carrying? (4) Is the problem written in the form of an equation? If the answers to the ten questions were the same for both the test item and the curriculum fragment we counted the item as covered.

Mathematics Concepts Subtest

The Mathematics Concepts subtest was more difficult to analyze because each item deals with a number of concepts at a fairly refined level. By examining publishers' outlines, which pointed out the skills that they were trying to emphasize, we developed criteria for matching the subskills required for each item. For example, in the measurement items we asked what basic operations students must use and whether the students were required to convert the measure into another unit of measure. In the geometry problems, our rules specified what shapes students were required to recognize, what geometric terms were used in the problem, and whether plane or solid geometry was required. With these levels of skills in mind, we searched the curriculum for exercises that would require only near transfer.

Mathematics Problem Solving Subtest

For the Mathematics Problem Solving subtest, we accepted the psychometric maxim that the reading level in the arithmetic story problems is deliberately set low so that arithmetic rather than reading is being tested. However, each arithmetic story item was examined and, if any item contained words of low frequency for a grade level (Carroll, Davies, and Richman, 1971), we eliminated it altogether from our analysis. The rules for deciding that the remaining arithmetic items had been taught thus included the features pertinent to arithmetic operations that we spoke of in the math computation rules, plus any specialized arithmetic words used in the problem.

Items That Could Not Be Analyzed

If we were unable to define clearly the skills needed for answering an item correctly, the item was dropped. For example, we eliminated one type of arithmetic story problem, the one labeled by test publishers as the multiple-step problem. The task requires several arithmetic operations and sometimes a conversion of measurements; steps can be performed in different orders to obtain a solution, and the sequence of skills is different for each pattern of solution. Because it was too difficult to create a list of skill combinations that would permit literal matching with curriculum units, these arithmetic problems were excluded from the evaluation.

The only other items eliminated from the Elementary and Advanced MAT battery were two items in the Word Knowledge subtest. In these items the target word was actually a combination of three or four words. For example, Item 19 on the Elementary battery was "A long wooden seat is a ...," and Item 50 was "Snow piled by wind is a ...." To get each of these items right, the student would have to encounter the same sequences of words in his PIP materials. Because we could not expect to find a curriculum exercise that would provide identical sequences, we dropped these items.

Procedures for Identifying Curriculum Materials that Match the MAT

A fair evaluation of the PIPs required that their posttest effectiveness be judged only on those MAT test items that students should have passed, given the PIP-specified curriculum materials that had been used. Our procedures for determining if the materials used in the fourth and eighth grades "taught" the skills required by the MAT items required a search through the curriculum materials that were specified in the PIP and used in the projects with children in those grades.

Another principle guiding our procedures was that we did not want to claim that students should have correctly responded to a particular item unless we could show with reasonable certainty that the curriculum materials contained lessons that would have "taught" the test item. Our rules for matching materials with test items were designed to be conservative, allowing only for near transfer. We wanted to be certain that materials meeting our criteria for corresponding with the MAT test items would permit the student to respond correctly to those items. We were not particularly worried about ignoring test items that had in fact been taught. But we were concerned about claiming that items had been taught when, in fact, they had not. Other professionals might be less stringent and might assume, for instance, that lessons in advanced vocabularly would guarantee that less advanced vocabulary was known. We did not assume this.

Dividing the Fourth and Eighth Grade PIP Specified-and-Used Curriculum Materials Into Units

To determine whether curriculum materials contained lessons or units that conformed to the rules we had established for each MAT item, we divided each material into "fragments." A fragment is a unit that deals with one skill. To identify appropriate fragments, we worked with both the SOIs and the materials. For each PIP-specified and used material that was analyzable, we looked at each student assignment and then examined the material to see how many distinct skills were covered in that assignment. If only one skill was covered, we adopted the system the teacher had used. If more than one skill was covered in the lesson, we divided it into fragments so that only one skill per fragment was covered.

For the math materials, we found it relatively easy to distinguish the various skills and separate the materials into fragments. The following examples are illustrative:

Material	Fragment	Example
Sullivan Mathematics	Book-page	Book 12, Page 3
Singer Individualized Math	Kit-block-lesson	Kit AA, Block 4, Lesson 12
Systems 80	Series-kit-lesson	Learning Number Facts, Kit B, Lesson 7

For the word knowledge materials, we had to be more arbitrary. Because we could not make each individual word a lesson, we depended more on the organization set by the publisher, as illustrated below:

Material	Fragment	Example
SRA Reading Labs	Labs-colors-lessons	Lab 1a, blue, Lesson 7
Systems 80	Series-kit-lesson	Reading Words in Context, Kit H, Lesson 3
Sullivan Programmed Reading	Book-page	Book 19, Page 90
McGraw-Hill Programmed Reading	Book-page	Book 20, Page 45
Random House Criterion Referenced Reading	Level/skill-lesson	46-13

Besides the series materials, there were individual books, like _Conquests in Reading_ and _Dr. Spello_, that included vocabulary words. They were generally broken down by page.

Like word knowledge materials, those for reading comprehension were broken into fragments suggested by the structure the publishers had created, but the fragments were generally larger because we made the assumption that reading is cumulative.

Material	Fragment	Example
Stories of the Inner City	Stories	Story 7
Barnell-Loft Specific Skills Series	Skill-book-lesson	Main Idea, Book A, Lesson 7
McCall-Crabbs	Book-lesson	Book A, Lesson 5
McGraw-Hill Programmed Reading	Book	Book 20
Random House Criterion Referenced Reading	Level/skill-lesson	48-5
Random House Reading Series	Level-difficulty	Orange, 5
SRA Reading Lab	Lab-color	Lab 1a, gold

Whenever SOIs failed to indicate how a material had been used, the site visitor asked the teacher. To be sure that materials were used as the publisher intended, we asked if tapes had been used with the reading materials, as expected in some kits, and if teachers had provided work sheets containing particular kinds of questions to test reading comprehension.

After determining that enough students had used a material, that we could obtain a copy, and that the material could be broken down into lessons, we were in a position to analyze the fourth and eighth grade curricula on a fragment-by-fragment basis. However, we did not analyze all material and all fragments within a material because this was beyond our resources. The next section describes the curriculum that we determined could be analyzed.

Curriculum Materials That Could Be Analyzed

Not all the materials specified by the PIP and used in the fourth and eighth grade classrooms were analyzed. Table G-2 shows which materials were not selected for analysis and the reasons for their rejection.

The least common of the five reasons for dropping a material were, first, that the material was not used by enough students to justify the time and cost of analyzing it or, second, that we were unable

Table G-2

ACCEPTABILITY OF PIP SPECIFIED-AND-USED MATERIALS FOR ANALYSIS

a. Catch-Up: Reading

Material	Word Knowledge: Analyzable	Word Knowledge: Not Analyzable	Reading Comprehension: Analyzable	Reading Comprehension: Not Analyzable
Scholastic Individualized Reading		4		4
Language Master		3		3
Criterion Reading, Random House	Yes		Yes	
Beginning to Read, Write, and Listen		1		1
Correlated Language Arts, Open Court		1		1
Reading Program, SRA		1		1
Reading Laboratory Kit--1a, SRA	Yes		Yes	
Reading Laboratory Kits 1b & 1c, SRA		1		1
Systems 80				
Concept Development		5		5
Learning Letter Sounds		5		5
Reading Words in Context	Yes			5
Barnell-Loft				
Getting the Facts		5	Yes	
Getting the Main Idea		5	Yes	
Drawing Conclusions		5	Yes	
Using the Context		5		5
Working with Sounds		5		5
Detecting the Sequence		5		5
Locating the Answer		5		5
Following Directions		5		5

b. Catch-Up: Math

Material	All Mathematics Subtests: Analyzable	All Mathematics Subtests: Not Analyzable
Sullivan Basal Mathematics	Yes	
Systems 80		
Learning Number Facts	Yes	
Developing Math Skills	Yes	
Singer Individualized Mathematics	Yes	
Tutor Computor		2
Criterion Reference, Random House	Yes	

Key: 1 = material used by too few students; 2 = material unavailable; 3 = indistinct lesson boundaries; 4 = inappropriate format for the MAT; 5 = inappropriate skills for the MAT.

Table G-2 (Continued)

c. Conquest: Reading

Material	Word Knowledge		Reading Comprehension	
	Analyzable	Not Analyzable	Analyzable	Not Analyzable
Basic Sight Vocabulary Cards, Dolch		3		3
Conquests in Reading	Yes			5
Coronet Cassettes & Workbooks		1		1
Dr. Spello	Yes			5
Controlled Reader		3		3
Merrill Phonics Skilltexts		5		5
Nicky	Yes		Yes	
McCall-Crabbs Standard Lessons in Reading		5	Yes	
Primary Reading Series, Bowmar		1		1
Programmed Reading, McGraw-Hill	Yes		Yes	
Phonovisual Wall Chart		5		5
Read, Study, Think		1		1
Reading Skill Builders, Reader's Digest		2		2
Systems 80				
Reading Words in Context	Yes			5
Learning Letter Sounds		5		5
Reading Laboratory Kit 1a, SRA	Yes		Yes	
Tachistoscope		3		3
Uncle Bunny		1		1
Xerox Microfilm Reader		3		3
Barnell-Loft				
Getting the Facts		5	Yes	
Using the Context		5		5
Locating the Answer		5		5
Working with Sounds		5		5
Following Directions		5		5

Key: 1 = material used by too few students; 2 = material unavailable; 3 = indistinct lesson boundaries; 4 = inappropriate format for the MAT; 5 = inappropriate skills for the MAT.

Table G-2 (Concluded)

d. HIT: Reading

Material	Word Knowledge		Reading Comprehension	
	Analyzable	Not Analyzable	Analyzable	Not Analyzable
Conquests in Reading	Yes			5
Remedial Reading Drills		5		5
Stories of the Inner City	Yes		Yes	
Sullivan Reading Program	Yes			5

e. HIT: Math

Material	All Mathematics Subtests	
	Analyzable	Not Analyzable
Ideal Flashcards for Addition		3
Self-Teaching Flashcards, Kenworthy		3
Sullivan Basal Mathematics	Yes	

f. IRIT: Reading

Material	Word Knowledge		Reading Comprehension	
	Analyzable	Not Analyzable	Analyzable	Not Analyzable
Mrs. Moon Series		5		4
Programmed Reading, McGraw-Hill	Yes		Yes	
Reading Laboratory Kit Ia, SRA	Yes		Yes	
Reading Program, Random House		5	Yes	
Barnell-Loft				
Getting the Facts		5	Yes	
Getting the Main idea		5	Yes	
Locating the Answer		5		5
Using the Context		5	-	5

Key. 1 = material used by too few students; 2 = material unavailable; 3 = indistinct lesson boundaries; 4 = inappropriate format for the MAT; 5 = inappropriate skills for the MAT.

to get a copy of the material to analyze. Eight of the materials were used by five students or fewer in all sites; for example, the Coronet Cassette and Workbook was used by only three children in one teacher's class. We were unable to obtain copies of two materials because the material was out-of-production (e.g., Tutor Computor), or the publisher would not lend a particular material that was too expensive to buy relative to the data it would provide.

More commonly, a material was not selected for analysis because it did not have distinct lesson boundaries. Materials such as flashcards, games, and teaching machines may cover (or not cover) a number of skills, depending on how they are used. For example, a student's schedule might show that he was assigned a lesson on the Language Master, but if the teacher did not note which skills she was working on in the lesson, we could not know which Language Master materials to include in our analysis. In addition, because teachers might elect to make their own cards to go with the Language Master machine, all we could know is that the child had some sort of audiovisual lesson.

The final two reasons for excluding a material from the analysis were that it had an inappropriate format for the MAT or that it had inappropriate skills. "Inappropriate format" meant that the student received help in reading the lesson; the MAT requires the student to read the items to himself. For example, the Mrs. Moon series has a tape for each book so the student can listen to the tape without reading the text. "Inappropriate skills" were apparent in various degrees. The match could have been as far off as "decoding skills" in the material and "word knowledge" skills on the MAT, or as close as reading comprehension paragraphs with the wrong kinds of questions. For example, the Specific Skills series by Barnell-Loft has reading comprehension passages, but some of its programs lack literal or inferential questions, having questions instead on sequence of events or on following directions. An example of a complete mismatch with the MAT is Remedial Reading Drills, which is used to emphasize phonics; the MAT Advanced has no items on phonics.

Our exclusion of certain materials from the analysis does not imply that those curriculum materials are bad or that the MAT tests the wrong skills. We were simply looking at what was PIP-specified and used in the classroom, and what parts of the MAT could be used to evaluate those particular programs.

Sampling the Reading Curriculum Materials

Except for reading comprehension materials, all PIP curricula that was specified, used, and analyzable (see Table 5-12) was examined for lessons that fit the rules we had written for the MAT items. In other words, the universe of the PIP curricula was searched for material that matched the items in the Elementary and Advanced Word Knowledge subtest and the three math subtests.

Because of the excessive number of reading comprehension materials and the cost of analyzing each, some sampling procedures were devised. The materials covered in projects where students took the reading subtests on the Elementary and Advanced levels of the MAT were selected for sampling. The sampling procedure for the most frequently used materials at the fourth grade is described below:

- SRA Reading Kit 1A--The kit is divided into finely graded sections, each in a different color and each containing 12 cards. The last card in each color, representing the most difficult in each section, was used as the sample.
- Random House Reading Series--Each of the two levels of books has 25 to 40 books, 10 to 50 pages in length. Within each level the books are graded and marked by difficulty (ten gradations). The sample for each book was 300 words drawn from the beginning, middle, and end of the book.
- McCall-Crabbs, Book A--This book has about 30 short reading passages, each aimed at slightly varying reading levels. Therefore, any passage, indicated in the SOIs as read, was analyzed.
- Random House Criterion Referenced Reading--This series is divided into five levels. Only the fourth level was analyzed; no pupil used the fifth level, and the passages in the third level were very short (35 words) compared with the shortest MAT passage (57 words on the Elementary). The fourth level contained about ten passages in each of two sections; five passages from each section were chosen randomly for analysis.
- Barnell-Loft--This series contains seven booklets for each skill: getting the facts, finding the main idea, and drawing conclusions. Within each skill, PIP

students read only the first five or six booklets, each of which contains 25 short passages. Where length was sufficient, four passages were chosen from each booklet--one from the beginning, two from the middle, and one from the end.

- McGraw-Hill Programmed Reading--This series has 21 booklets, each of 124 pages. Only Books 8-21 were analyzed; below Book 8 the passages consist of two or three short sentences. In each booklet three or four tests are given to check mastery of preceding content. The longest story preceding each test was used as the sample. When the story was too long (10 pages), 200 words from the middle of the story were sampled. This material, then, was sampled most heavily because it is by far the longest series and because it was frequently used in several PIPs.

The only analyzable reading material used in the eighth grade was a 200-page book, divided into approximately 20 stories, each of which was 900 to 3000 words long. A sample of 225 words (nearest complete sentence) was drawn from the middle of each story.

Results of the Curriculum Search for Matches with MAT Elementary and Advanced

Tables G-3 and G-4 display the results of matching the descriptions of the MAT items with the PIP materials. Two general comments can be made about the results of this analysis. First, the skills needed for both the Reading and the Word Knowledge subtests were generally covered somewhere in the curriculum, although we found that all PIP curricula analyzed placed heavy emphasis on phonics, decoding, and word attack skills. These skills are not included in the MAT tests above Primer. Second, the PIP curricula concentrated on basic operations presented in the "old math" style; a few concepts were discussed if they related to basic operations. The MAT divides its items into "old" and "new" math and into basic and more advanced math. The MAT also places heavy emphasis on understanding mathematical concepts.

- Elementary Word Knowledge--Table G-3a shows the words used in the first three items of the Elementary Word Knowledge subtest. To ensure test security, the remaining 4 items shown in the table include only the words "target," "context," and "answer."

Table G-3

CORRESPONDENCE BETWEEN ELEMENTARY MAT ITEMS
AND SPECIFIED-AND-USED MATERIALS FOR ALL PIPS

a. Word Knowledge Subtest

MAT Item Number	Conquests in Reading (number of entries)	Dr. Spello (number of entries)	*McGraw-Hill* (book number)	Nicky (number of entries)	Random House (number of entries)	SRA (number of entries)	Systems 80 (number of entries)	Full Item
1. Night is the opposite of day	1	1 3 3	12 15-21 7-20	1	1 3	5 1	1	X
2. Meat is a type of food			15-17,20,21 16-21		1 1	3	1	X
3. Happy means glad	7 1	1 2	9-17 15-21 4,6-9,11,15,16,18	1	2 2 1	1 7 2	1	X
4. Target Context Answer	1 1	2	6-14	1	4	1 13 1	1 1	X
5. Target Context Answer	1	1	16,17,20 13,14,17-19,21 11,13-21		2	1	1	X
6. Target Context Answer	1 3	2	12,14-21 15-21 12-14,16,17,19	1	2	4		X
7. Context Target Answer	1 1 1	3	15-21 14 8-10,12-14,17	1 1	3 3 3		1	X
8. Target Context Answer	1	2	13-19 5,6,9,11,12		1	4		
9. Context Target Answer	1	3 1	15-21 15,16 13-17,19,21	1	3 1	1		X
10. Target Context Answer		2	13-19 17,20		1			
11. Target Answer		0	6-8,16-18				1	
12.[a] Target Answer			14					
13. Target Answer			16,19 16-21			1	1	X
14. Target Answer			2-6,15,19			3		
15. Target Context Answer		2	13-19 8,10,12-21		1			
16. Target Context Answer	1	1 1	17,19-21 9-21 12-14,19			14	1	X
17. Target Answer			10,12 18,20	1		1		X
18. Target Answer			15-21 19			3		X
19.[a]								
20. Target Context Answer	1	8	8-21 10,13-20		1	1	1 1	
21.[a]								
22. Target Context Answer	1 7	2	10-17,19-21 15-21 18,19	1	2	1 7		X

*Book numbers listed indicate the books in which the word is covered six times or more.

[a] No rules could be developed for this item.

Table G-3 (Continued)

a. Word Knowledge Subtest (Continued)

MAT Item Number	Conquests in Reading (number of entries)	Dr. Spello (number of entries)	McGraw-Hill* (book number)	Nicky (number of entries)	Random House (number of entries)	SRA (number of entries)	Systems 80 (number of entries)	Full Item
23. Target Answer			17-19,21 21			1		X
24. Target Context Answer	7	2	15-18,20,21 15-21	1	2	7		
25. †								
26. †								
27. Target Answer			9,12,16,20,21					
28. Target Context Answer			20,21 7-16 11,20		2	2	1	X
29. †								
30. Target Answer			19-21 9-14,16,17,19,20			5		X
31. Target Answer						2 1		X
32. Target Answer			15,16					
33. Target Answer			11					
34. Target Context Answer	7	2	16-21 15-21 6-8,14,15	1 1 1		7		X
35. Target Context Answer	6	12	20 13-19 8,9,15,17,21		1	3	3	X
36. Target Answer								
37. Target Context Answer	7	2 1	20,21 15-21 12,14,18,20	1	[illegible]	1 7		X
38. Target Answer		1	17-21 7,9,11,13,19,21			6		X
39. Target Answer			6,13-15			1 3		X
40. Target Answer			16,19			1		
41. Target Answer			17,20,21 21					X
42. Target Context Answer	7	1 2	8,10 15-21 19,20	1	2	7 1	1	X
43. Target Answer			12-17,20,21					
44. Target Context Answer		2	21 15-21	1	2	7		
45. Target Context Answer	7	2	16 15-21 16,18-21	1	2	7 1		X
46. Target Answer			17,20,21 20,21			1		X
47. Target Context Answer	7	2	20,21 15-21 12-16,20,21	1	2	7 1		X

*Book numbers listed indicate the books in which the word is covered six times or more

†No rules could be developed for this item.

Table G-3 (Continued)

a. Word Knowledge Subtest (Concluded)

MAT Item Number	Conquests in Reading* (number of entries)	Dr. Spello (number of entries)	McGraw-Hill* (book number)	Nicky (number of entries)	Random House (number of entries)	SRA (number of entries)	Systems 80 (number of entries)	Full Item
48. Target Answer			15					
49. Target Answer	1		21 20,21					X
50.[†]								

b. Reading Subtest

MAT Item Number	QT	PL	SL	Spache	Spache/DC	Random House Reading: Blue Level	Random House Reading: Orange Level	SRA Lab Ia	McCall-Crabbs Book A (lesson)	Random House Criterion Reference Level 4 Skill 8 (lesson)	Barnell-Loft Drawing Conclusions (book)	Barnell-Loft Getting the Facts (book)	Barnell-Loft Getting the Main Idea (book)	McGraw-Hill Programmed Reading (book)
01	3	56	11.2	3.9	3.9		5							
02	4						5					E		20,21
03	3						5				D			
04	3						5							20,21
05	1	[illegible]	10.6	3.2	3.2							E		20,21
06	[illegible]									3			D	
07	4					10	4		3,5,7,14			D,E		14-21
08	4					10	4		5,7	1,4	D			
09	3					10	4		5,7	1,4	D			
10	4					10	4		3,5,7,14			D,E		14-21
11	3	65	9.2	2.5	2.5	10	4		5,7	1,4	D			
12	4					5,10	4	Blue	3,5-7,9,10,14			A-E		10-21
13	2					5,10	4	Blue	5-7,9,10	1,2,4	D			
14	4													
15	4					5,10	4	Blue	5-7,9,10	1,2,4	D			
16	4					5,10	4	Blue	5-7,9,10	1,2,4	D			
17	1	65	10.8	4.0	4.2	5,10	4	Blue	5-7,9,10	1,2,4	D			
18	3													
19	2													21
20	4													
[illegible]	3	82	13.6	4.2	4.6						E			
[illegible]	4													
[illegible]	3													
[illegible]	[illegible]													
25	[illegible]													
26														
27														
28	4													
29	[illegible]	[illegible]	11.0	4.7	5.8									
30	4													
31	4													
32	3													
33	2													
34	3													
35	3	[illegible]	14.8	4.4	5.7									
36	4													
37	2													
38	3													
39	4													
40	1	79	13.1	5.3	6.4									
41	3													
42	4													
43	4													
44	4													
45	3													

QT = question type: 1 = main idea, 2 = word in context, 3 = literal, 4 = inferential; PL = passage length, SL = sentence length; Spache/DC = Average Spache and Dale-Chall.

*Book numbers listed indicate the books in which the word is covered six times or more.

†No rules could be developed for this item.

Table G-3 (Concluded)

c. Mathematics Subtests

	Math Computation				Math Concepts						Math Problem Solvin		
MAT Item Number	Sullivan Basal Mathematics (book)	Systems 80 Learning Number Facts (kit)	Singer Kit AA (block)	Singer Kit BB (block)	Sullivan Basal Mathematics (book)	Systems 80 Developing Math Skill (kit)	Systems 80 Preschool (kit)	Random House Criterion Reference (level)	Singer Kit AA (block)	Singer Kit BB (block)	Sullivan Basal Mathematics (book)	Singer Kit AA (block)	Singer Kit BB (block)
01	15	B	1,3				B5				13	4,8	
02	3	A	1,3								8	1,8	
03	13		1,3								16	4,8	
04	16		1,3								14	4,8	
05	16		6,8										
06	16												
07	18		5										
08	18		5		36				9	6			
09	17	D	2,3										
10	8	C	2,3								36		
11	16		6,8										
12	16		6,8								*	*	*
13	36												
14	25	E		7									
15	18		8										
16	36												
17	18		8										
18	[illegible]												
19	24	E											
20	25				36				9	6			
21	36					AA			9	6			
22	24												
23	27	G						3,4			*	*	*
24	29	G									*	*	*
25	28												
26	27	G			36				9	6			
27	36												
28	18		8										
29	31												
30	27	G											
31	24										*	*	*
32	25												
33	25												
34	32												
35	19		8						9	6			
36	29												
37	31												
38	27												
39	33												
40	22												

*No rules could be developed for this item.

Table G-4

CORRESPONDENCE BETWEEN ADVANCED MAT ITEMS
AND SPECIFIED-AND-USED MATERIALS FOR ALL PIPS

a. Word Knowledge Subtest

MAT Item Number	Sullivan (BRL) Book Number	Sullivan (BRL) Number of Entries	Stories of Inner City (number of entries)	Conquests in Reading (number of entries)	Full Item
1. Demolished means destroyed	8	1	1	7 1	X
2. Used is the opposite of new	5	1		1 1	
3. To comment is to remark					
4. Target Context Answer	4 1	1 1		1	
5. Target Context Answer	8 10	1 1		1 7 2	X
6. Target Context Answer	6	1			
7. Target Context Answer	8	1		7	
8. Target Answer					
9. Target Answer					
10. Target Answer					
11. Target Answer			1		
12. Target Answer	18	1			
13. Target Answer					

401

Table G-4 (Continued)

a. Word Knowledge Subtest (Continued)

MAT Item Number	Sullivan (SRL) Book Number	Sullivan (SRL) Number of Entries	Stories of Inner City (number of entries)	Conquests in Reading (number of entries)	Full Item
14. Target					
Answer					
15. Target					
Context		1		7	
Answer				1	
16. Target					
Context					
Answer	6	1			
17. Target					
Context	6	1			
Answer	3,13	2			
18. Target					
Answer	10	1			
19. Target					
Context	6	1			
Answer					
20. Target					
Context					
Answer					
21. Target					
Answer					
22. Target					
Answer					
23. Target					
Answer					
24. Target				1	
Answer					
25. Target					
Answer					
26. Target					
Context	8	1			
Answer	11	1			
27. Target					
Answer	8	1			

Tabld G-4 (Continued)

a. Word Knowledge Subtest (Continued)

MAT Item Number	Sullivan (BRL) Book Number	Sullivan (BRL) Number of Entries	Stories of Inner City (number of entries)	Conquests in Reading (number of entries)	Full Item
28. Target					
Context	8	1			
Answer	25	1			
29. Target					
Answer					
30. Target					
Context	8	1			
Answer					
31. Target					
Context	8	1			
Answer				1	
32. Target			1		
Context	8	1			
Answer					
33. Target					
Answer					
34. Target					
Answer				1	
35. Target					
Answer					
36. Target					
Context	8	1			
Answer	6	1			
37. Target					
Context					
Answer	12	1		1	
38. Target					
Answer			1		
39. Target					
Context	8	1			
Answer					
40. Target					
Answer	25	1	1		
41. Target					
Answer			1		

Tabld G-4 (Continued)

a. Word Knowledge Subtest (Concluded)

MAT Item Number	Sullivan (BRL) Book Number	Sullivan (BRL) Number of Entries	Stories of Inner City (number of entries)	Conquests in Reading (number of entries)	Full Item
42. Target					
Context	17	1		1	
Answer					
43. Target					
Answer					
44. Target					
Context	5	1			
Answer	1	1			
45. Target					
Context	8	1			
Answer					
46. Target					
Context	8	1			
Answer					
47. Target			1		
Context	8	1			
Answer					
48. Target				1	
Answer					
49. Target				2	
Answer					
50. Target					
Context	8	1			
Answer					

Table G-4 (Continued)

b. Reading Subtest

MAT Item Number	Question Type*	Passage Length	Sentence Length	Dale-Chall	Stories of Inner City (pages)
01	4	104	17.3	8.8	
02	2				
03	3				
04	1				
05	4	148	16.4	7.5	214-218
06	1				
07	2				
08	4				214-218
09	4				214-218
10	4	205	15.6	8.1	
11	4				
12	4				
13	4				
14	2				
15	4				
16	4	159	15.9	10.[illegible]	
17	3				
18	3				
19	4				
20	2				
21	2				
22	1				
23	4	260	18.5	11.5	
24	4				
25	4				
26	3				
27	3				
28	2				
29	2				
30	2	302	18.8	9.1	
31	4				
32	1				
33	3				
34	4				
35	2				
36	2				
37	3				
38	4	316	15.0	10.5	
39	4				
40	4				
41	4				
42	4				
43	4				
44	2				
45	4				

*1 = main idea, 2 = word in context, 3 = literal, 4 = inferential.

Table G-4 (Concluded)

c. Mathematics Subtests

Math Computation		Math Concepts		Math Problem Solving	
MAT Item Number	Sullivan Basal Mathematics (book)	MAT Item Number	Sullivan Basal Mathematics (book)	MAT Item Number	Sullivan Basal Mathematics (book)
01	24	01	36	01	24
02	11	02	*	02	28
03	27	03		03	*
04	21	04	31	04	*
05	31	05		05	
06	16	06		06	36
07	20	07		07	
08	18	08		08	
09	36	09		09	*
10	29	10	*	10	
11	31	11		11	*
12	25	12		12	32
13		13		13	
14	27	14	*	14	
15	36	15		15	
16	31	16		16	
17		17		17	
18	36	18		18	*
19	32	19		19	
20	36	20	36	20	*
21		21		21	
22	32	22		22	
23	25	23		23	
24	32	24		24	36
25		25		25	
26	*	26		26	
27	32	27		27	
28	32	28		28	
29	36	29		29	
30	32	30		30	*
31	36	31		31	
32		32		32	
33	36	33		33	
34		34		34	
35		35		35	
36	33	36			
37		37			
38		38			
39		39			
40		40			

*No rules could be developed for this item.

Some of the items do not list the context word, which means that the context word was "is" or "is to." We reviewed the elementary literature and found that "is" was so commonly used that it was not necessary to check it for each individual item. The numbers under each material show the number of times that a particular word was found in that material. Book numbers are shown for the McGraw-Hill series, however, so that the reader may see how the words are distributed across the 21 books.*

While many of the MAT words were found in the materials, the data base for our finest level of analysis was restricted to MAT items for which each of the three words were found. These items are labeled "Full Item" in the right-hand column. Relative to Item 8, for example, the word "tent" was found, but the other two (teepee, kind) were not. Although about 87% of the MAT words were found in the materials, only 56% of the found words were part of full items.

- Elementary Reading--Table G-3b shows the correspondence between the Elementary Reading subtest and the materials. The question type (QT) and passage features for each story are listed for each item.

The beginning levels of the materials [SRA, Random House Criterion Reading, McGraw-Hill, and Barnell-Loft (Getting the Main Idea and Drawing Conclusions)] did not contain passages of sufficient length to meet the requirements of the test's readability analysis (vocabulary, mean sentence length, and passage length). Few materials lacked the requisite types of questions, with the notable exception of word-in-context. The PIP fourth and eighth grade materials emphasized literal and inferential questions.

Since the Elementary Reading subtest was given at pre- and post-test to fourth graders, the passage features of the curriculum materials cluster about the items at or below the fourth grade level. Thus, these materials do not match about 60% of the test items.

---

*The book numbers listed on Table G-3a indicate the books in which a particular word is covered six times or more.

- Elementary Mathematics Computation--Table G-3c shows the correspondence between the materials and all three of the Elementary math subtests. An asterisk indicates the items for which we could not develop rules. All items in the Elementary Mathematics Computation subtest were found in the curriculum materials. Basic operations are emphasized in the test and in Catch-Up.

- Elementary Mathematics Concepts--While the materials covered approximately 18% of the items on this subtest, congruence is spotty and limited to part of a Sullivan booklet, one Systems 80 card, and a few pages in a reading series.

- Elementary Mathematics Problem Solving--Few of the PIP materials in Catch-Up contain story problems like those in this subtest. The Sullivan series contains none at all. We considered only the most simple problems (basic operations) to be candidates for correct responses.

- Advanced Word Knowledge--Table G-4a lists the word knowledge items for the Advanced MAT and the corresponding lessons in the PIP materials. The number of MAT items found in the materials were far fewer than for the Elementary test. While this difference can be partly attributed to the smaller number of materials in HIT, the difference is probably also due to difficulties in sampling the larger vocabulary of eighth graders. Even for the two items covered in toto, the target word for Item 1 was found on only one page and the correct response on only one page. Approximately 50% of the words--either target, context, or answer--were not found in the materials.

- Advanced Reading--Table G-4b shows that only one material, Stories of the Inner City, contained questions and passages similar to those on the test. While many stories had the appropriate questions, the passage feature indices were considerably lower than those of 90% of the test items.

- Advanced Mathematics Computation--The only material analyzed for math subtests was the Sullivan Basal Mathematics series (see Table G-4c). As in the fourth grade, basic operations were emphasized in the eighth grade, but the Advanced MAT included more higher-level math problems. This series matched about 70% of the test items.

- Advanced Mathematics Concepts--The Sullivan series contains few concepts other than those pertinent to basic operations on whole numbers, fractions, and decimals. The Mathematics Concepts subtest consists of many items (about 95%) that cover skills, now labeled new math, that are not covered in the eighth grade curriculum.

- Advanced Mathematics Problem Solving--The Sullivan series does not contain any story problems similar to those in this subtest. Consequently, only the most simple problems (basic operations) were considered to be candidates for learning due to PIP instruction.

In reviewing this comparison of the MAT and the PIP-specified and used curriculum, the reader is reminded that we were able to look only at the materials known to be covered from January to posttesting. We know nothing about assignments from pretesting to January, nor do we have concrete information about lessons assigned in the regular curriculum. Although we are confident that the SOIs reported most of the PIP treatment for students in the fourth and eighth grades from January to posttesting, we cannot guarantee the inclusion of all lessons assigned. Some of the MAT items we have excluded from the analysis might be relevant to materials we have not examined. Possibly the MAT is more relevant to the fourth and eighth grade PIP curricula than it appears from our analysis, but we are limited to conclusions that can be drawn from the assignments reported on the SOIs.

Table G-5 displays those MAT item numbers for each subtest in the Elementary and Advanced battery that were covered by at least one fragment used in the fourth or eighth grade. "N" represents the greatest number of items in our analysis any student could have answered correctly, if he covered all of the PIP-specified materials used in the projects. Only in Math Computation does the MAT appear relevant to the fourth and eighth grade curriculum. That the projects with a mathematics component do well on this MAT subtest is probably because of the greater relevance of this part of the MAT to the projects' objectives.

Attempts to Match Each Student with the Items He Had Covered

As noted earlier, we planned to analyze the relevance of the MAT to the curriculum at two levels. We have described the intermediate level in the previous sections. In this section we describe the more detailed level, where we matched each student with the items he covered.

Tabld G-5

THE MAT ITEMS KNOWN TO BE COVERED BY PIP-SPECIFIED MATERIALS

a. Elementary Battery

Word Knowledge	Reading	Mathematics: Computation	Mathematics: Concepts	Mathematics: Problem Solving
1	1	1	1	1
2	2	2	8	2
3	3	3	20	3
4	4	4	21	4
5	5	5	23	10
6	6	6	26	25
7	7	7	35	T = 35
9	8	8	T = 40	N = 6
13	9	9	N = 7	N = 17.1% of T
16	10	10	N = 17.5% of T	
17	11	11		
18	12	12		
22	14	13		
23	15	14		
28	16	15		
30	18	16		
31	20	17		
34	T = 45	18		
35	N = 17	19		
37	N = 37.7% of T	20		
38		21		
39		22		
41		23		
42		24		
45		25		
46		26		
47		27		
49		28		
T = 50		29		
N = 28		30		
N = 56% of T		31		
		32		
		33		
		34		
		35		
		36		
		37		
		38		
		39		
		40		
		T = 40		
		N = 40		
		N = 100% of T		

Key: T = total number of items in MAT subtest; N = number of MAT items known to be covered by PIP-specified materials.

Tabld G-5 (Concluded)

b. Advanced Battery

Word Knowledge	Reading	Mathematics: Computation	Mathematics: Concepts	Mathematics: Problem Solving
1	5	1	1	1
5	8	2	4	2
T = 50	9	3	20	6
N = 2	T = 45	4	T = 40	12
N = 4% of T	N = 3	5	N = 3	24
	N = 6.6% of T	6	N = 7.5% of T	T = 35
		7		N = 5
		8		N = 14.2% of T
		9		
		10		
		11		
		12		
		14		
		15		
		16		
		18		
		19		
		20		
		22		
		23		
		24		
		27		
		28		
		29		
		30		
		31		
		33		
		36		
		T = 40		
		N = 28		
		N = 70% of T		

Key: T = total number of items in MAT subtest; N = number of MAT items known to be covered by PIP-specified materials.

The methodology for the detailed analysis required returning to the SOIs and recording each student's PIP-specified assignments. We did this for students who took either the Elementary or the Advanced MAT and whose teachers had either a high or low (5, 6, 8, or 9) implementation rating. Exhibit G-2 displays the form used to inventory the PIP-specified assignments for a student in Conquest. These inventory forms included all materials analyzed in Section 5.7.

The next three steps (already completed in the intermediate-level analysis) were: (1) analyzing the Elementary and the Advanced MAT items for the skills needed for the appropriate answers (described in Section 5.7), (2) analyzing the curriculum to find the lessons that taught those skills, and (3) developing a list (Tables G-3 and G-4) of the MAT items that had been covered by some part of the PIP specified-and-used curriculum.

The final step required determining the correspondence between each student's inventory of PIP assignments and the MAT items covered by those assignments. This was done by overlaying each student's inventory of PIP assignments on the list of MAT items covered by the curriculum materials (Tables G-3 and G-4). To determine whether a student had passed the items he had covered (that he should have answered correctly), we planned to sort each student's file as illustrated by the following schematic:

	MAT Items									
	1	2	3	4	5	6	7	8	9	10
Student's pretest	0	0	0	0	1	1	1	1	1	0
Covered by PIP curriculum	0	1	0	1	0	1	0	1	0	1
Student's posttest	0	1	1	0	0	0	1	1	1	1

A "0" or a "1" for the pre- or the post-test indicates a child's failure or success, respectively, to correctly answer a particular MAT item; a "0" or a "1" for the "covered by PIP curriculum" indicates the student's lack of exposure or his exposure to materials related to that MAT item.

From the outset we anticipated having too few data points (students and materials/MAT correspondences) to warrant completing the final sort into correct-incorrect responses or performing the analysis. After we had determined which items each student had covered, it became apparent that there were indeed two few data points.

Table G-6 displays the results of the last step we completed. The number of students who had covered at least one MAT item in their curricula

Exhibit G-2

INVENTORY OF INDIVIDUAL STUDENT'S PIP-SPECIFIED ASSIGNMENTS

Conquest—Reading
Site — Cleveland
Page 3/4

Teacher Jane Doe
Grade 4
Rating 1

	Ronald Hooks	Paul Harris		
Barnell-Loft, Specific Skills (cont.)				
Using the Context	B, C			
Locating the Answer				
Working with Sounds	B, C			
Following Directions				
Detecting the Sequence	B			
Webster Classroom Clinic — Dr. Spair	✓			
Lesson #30 (page 18) " 32 (" 19) " 33 (" 20) " 47 (" 25) " 48 (" 28) " 49 (" 29) " 50 (" 29) " 52 (" 30) " 53 (" 31) " 54 (" 31) " 55 (" 32) " 56 (" 32) " 57 (" 33) " 140 (" 94)				
Read, Study, Think	@			
Pamphlet 1 (Page #) 2 3 4 5 6	@, p. 9, 12, 14, 18			
Merrill Phonics Skill Text				
Book A (Page #) B C D E F				

Conquest—Reading
Site — Cleveland
Page 4/4

Teacher Jane Doe
Grade 4
Rating 1

	Ronald Hooks	Paul Harris		
Nicky				
Uncle Bunny				
Individual Skill Builders Readers Digest	✓ a lot			
Phonovisual Wall Chart				
Bowmar Primary Reading Series Primary Level				
McGraw Hill Programmed Reading	B12 p. 3-21, 56-130 B13 p. 23-30			
Getting the Main Idea Book A B C	✓			
Book D unit 1 11 25				
Book E unit 2 10 24				
Book F unit 25				

Table G-6

NUMBER OF ITEMS COVERED BY STUDENTS
IN EACH PIP: GRADE 4 or GRADE 8

	Elementary MAT			Advanced MAT
	Catch-Up	Conquest	IRIT	HIT
(A) Number of students in grades 4 or 8 whose teachers were rated as good or bad	27	30	28	63/40*
Total Reading				
(B) Number of students in (A) who covered one or more items on the MAT in their individual reading curricula	8	25	23	1
(C) Number of reading items possible for students in (B)+	360	1125	1035	5
(D) Number of reading items covered by students in (B)	9	129	227	3
Total Math				
(E) Number of students in (A) who covered one or more items on the MAT in their individual math curricula	21	NA	NA	34
(F) Number of math items possible for students in (E)	1013	NA	NA	1178
(G) Number of math items covered by students in (E)	189	NA	NA	162

NA = Not applicable.

*There were 63 students in the HIT reading program and 40 students in the HIT math program.

+The number of possible items for all students in Group (B) was calculated by multiplying the number of possible items (Table 5-15) on each subtest by the number of students with valid scores on that subtest.

was quite small--much smaller than the number being considered in our fourth and eighth grade sample. Moreover, the number of items actually covered, when compared with the number of possible items, was again quite small. This was especially obvious in light of our plans, which called for analyzing data within project by comparing the scores of children who had well-implemented/responsive teachers with the scores of children who had poorly implemented/nonresponsive teachers. With so few items, there did not appear to be enough data in each category to analyze effects.

REFERENCES

Beck, M. D., "Development of Empirical 'Growth Expectancies' for the Metropolitan Achievement Tests," paper presented at the 1975 convention of the National Council on Measurement in Education, Washington, D.C., 31 March 1975.

Biomedical Computer Programs, Health Sciences Computing Facility, Department of Biomathematics, School of Medicine, UCLA, Los Angeles, California (University of California Press, Berkeley, California, June 1973).

Calfee, R. C., "Perceptual and Memorial Components in Beginning Reading," interim report to Carnegie Corporation (January 1975).

Carroll, J. B., P. Davies, and B. Richman, The American Heritage: Word Frequency Book (Houghton Mifflin Co., Boston, Massachusetts, 1971).

Coleman, J. S., et al., Equality of Educational Opportunity, U.S. Department of Health, Education, and Welfare, Office of Education (Government Printing Office, Washington, D.C., 1966).

Dale, E. and J. S. Chall, "A Formula for Producing Readability," Educational Research Bulletin, Vol. 27, pp. 11 ff. (21 January 1948) and Vol. 27, pp. 37-54 (18 February 1948).

Golub, H., and C. Kidder, "Syntactic Density and the Computer," Elementary English, Vol. 51, No. 8, pp. 1128-1131 (November/December 1974).

Hoepfner, R., et al., "CSE Elementary School Test Evaluations," Center for the Study of Evaluation, Graduate School of Education, UCLA, Los Angeles, California (1970).

Horst, D. P., G. Tallmadge, and C. Wood, "A Practical Guide to Measuring Project Impact on Student Achievement," U.S. Department of Health, Education, and Welfare, Office of Education, Washington, D.C. (1975).

Hunt, K., "Grammatical Structures Written at Three Grade Levels," NCTE Research Report No. 3, National Council, Teachers of English (1965).

Jencks, C., Inequality: A Reassessment of the Effect of Family and Schooling in America (Basic Books, Inc., New York, New York, 1972).

Judd, W., "Computer Program for Spache and Dale-Chall Readability Formulae," 322 College Avenue, No. D, Palo Alto, California (1975).

Kaskowitz, D., and C. Norwood, "A Study of the Norm-Referenced Procedure as Applied to the Evaluation of Project Information Packages," Stanford Research Institute, Menlo Park, California (to be published, 1976).

Klare, G., "Assessing Readability," Reading Research Quarterly, Vol. X, No. 1, pp. 62-102 (1974-75).

Lord, F. M., and M. R. Novick, Statistical Theories of Mental Scores (Addison-Wesley Publishing Co., Inc., Reading, Massachusetts, 1968).

Loret, P. G., et al., Anchor Test Study: Equivalence and Norm Tables for Selected Reading Achievement Tests (Grades 4, 5, 6) (Government Printing Office, Washington, D.C., 1974).

MAT Guidelines No. 1, Development of the Standard Score System for the 1970 Edition of Metropolitan Achievement Tests (Harcourt Brace Jovanovich, Inc., New York, New York, October 1972).

Nie. N. H., et al., Statistical Package for the Social Sciences (McGraw-Hill Book Co., New York, New York, 1970).

Pelavin, S., and P. Barker, "A Study of the Generalizability of the Results of a Standardized Achievement Test," based on the Rand Corporation study for the National Institute of Education under Contract B2C-5326, paper presented at the American Educational Research Association Meeting, San Francisco, California, 19-23 April 1976.

Piestrup, A. M., "Design Considerations for Packaging Effective Approaches in Compensatory Education," Technical Report UR-241, RMC Research Corporation, Mountain View, California (1974).

Spache, G., "A New Readability Formula for Primary Materials," in-house report, University College, University of Florida, Gainsville, Florida (unpublished).

Tallmadge, G. K., "The Development of Project Information Packages for Effective Approaches in Compensatory Education," ~~Technical Report~~ UR-254, RMC Research Corporation, Mountain View, California (1974).

Tuinman, J., "Determining the Passage Dependency of Comprehension Questions in Five Major Tests," Reading Research Quarterly, Vol. IX, No. 2, pp. 206-223 (1973-74).

Wargo, M. J. and D. R. Green, "Achievement Testing of Disadvantaged and Minority Students for Educational Program Evaluation," CTB McGraw Hill (in press).