DOCUMENT RESUME

ED 112 635 FL 005 470

AUTHOR Seelye, H. Ned; Balasubramonian, K.
TITLE Accountability in Educational Reform Programs through Instrumentation Analyses and Design Variation: Evaluating Cognitive Growth in Illinois Bilingual Programs, 1972-73.
PUB DATE Feb 73
NOTE 30p.

EDRS PRICE MF-$0.76 HC-$1.95 Plus Postage
DESCRIPTORS *Bilingual Education; Bilingualism; Cognitive Measurement; *Educational Accountability; Educational Objectives; Elementary Education; Evaluation Criteria; *Evaluation Methods; Evaluation Needs; Measurement Instrument S; Program Design; *Program Evaluation; Spanish Speaking; *State Aid; Student Evaluation
IDENTIFIERS Elementary Secondary Education Act Title VII; ESEA Title VII; *Illinois

ABSTRACT

The bilingual situation in Illinois is described briefly, and an outline of the instructional objectives of local bilingual programs is given. The programs are to be: (1) measurable and oriented toward the end-of-year-product, and (2) organized within the guidelines for state-funded bilingual programs. The main part of the report describes the design of the procedures set up to evaluate these programs based on the following recommendations from the Office of the Superintendent of Public Instruction: (1) prior to implementing a bilingual program in a community a sociolinguistic survey should be conducted there; (2) priority should be given to early childhood programs, preferably pre-school and kindergarten; (3) 'standardized' instruments, rather than criterion-referenced tests should be selected as measurement tools; and (4) insofar as possible, a true experimental evaluation design should be employed, with randomly assigned treatment and control groups. The aim was to select and implement the combination of designs and instruments which would most effectively give an accurate picture of local bilingual education programs. Actual evaluation findings are not reported here. Anticipated design refinements for future years are mentioned, and three tables give: (1) a description of the measuring instruments; (2) statewide evaluation designs and project sites, and (3) between-groups hypothesis. (TL)

# ACCOUNTABILITY IN EDUCATIONAL REFORM PROGRAMS THROUGH INSTRUMENTATION ANALYSES AND DESIGN VARIATION: EVALUATING COGNITIVE GROWTH IN ILLINOIS BILINGUAL PROGRAMS, 1972-73

H. Ned Seelye
Office of the Superintendent of Public Instruction
Chicago, Illinois

K. Balasubramonian
Bilingual Education Service Center
Mount Prospect, Illinois

Unlike most other states with large non-English-speaking populations, most Illinois bilingual programs are funded from state revenues. In the short span of three years, state funds for bilingual education have increased dramatically from $200,000 to $2,370,000. At this writing (February, 1973), forty-nine bilingual programs are state funded, nine are federally funded (ESEA Title VII), and one is funded by the Chicago Board of Education. (The city of Chicago also contributes to some of the other bilingual programs above the city-wide per capita expenditure level.) Twenty-eight of the fifty-nine bilingual programs are outside the city of Chicago. Most of these "downstate" programs fall within the wide geographic band which stretches west to Moline on the Iowa border, north to Waukegan and Rockford near the Wisconsin border, and south to Joliet. A few programs go as far south as Danville and Arcola.

---

NOTE: Since this paper was written, the Illinois General Assembly appropriated $6,000,000 for bilingual programs in FY-74. This additional revenue allowed the number of Chicago projects to increase to 57, and the downstate projects to 35. The number of children served in bilingual programs jumped from 5,000 to 16,000.

Between the two-thirds and three-fourth of the children who need a bilingual program live in Chicago. Headcounts have identified 65,000 of these children in the Chicago Spanish-speaking community alone. Schools need help as they try to meet the special educational needs of children who, because they understand another language and have learned the values of another culture, will not approach their own potential for learning in our traditional English-language curriculum. Of the estimated 100,000 Illinois children from non-English-speaking backgrounds, less than six percent are currently enrolled in a bilingual program.

The instructional objectives of bilingual programs are developed by each project to suit their local needs. This is accomplished within the parameters of two constraints; the objectives are to be measurable, end-of-year product oriented, and they are to be organized under the appropriate goal described in the state guidelines for all bilingual programs seeking state reimbursement. There are seven of these goals.

(1) Children in the bilingual program will achieve fluency and literacy in two languages.

(2) Children in the bilingual program will achieve at a rate commensurate with their own age, ability, and grade level in all school subject areas.

(3) Children in the bilingual program will demonstrate growth in self-esteem.

(4) Children in the bilingual program will be provided with a coordinated and integrated learning environment through effective coordination with the regular school program.

(5) All teachers and staff members of participating schools will be involved in a comprehensive inservice training program.

(6) Parents and other community members will be involved in the planning, implementation, and evaluation of the bilingual program.

(7) Each bilingual project will implement an evaluation to assess its effectiveness.

Much of the negative findings reported by recent studies of compensatory educational programs and experiments in performance contracting (e.g., Garfinkel, 1972) has been criticized as chronologically premature and analytically faulty (Campbell and Erlebacher, 1970; Campbell and Frey, 1970; O'Connor and Klein, 1972). The critics underscore the need for alternate procedures in data analysis and interpretation. Wrightstone (n.d.) and Fitzgibbon (n.d.) outline a number of cautions and suggest preferable procedures to be employed in measurement tasks, especially in the use of standardized tests for the purposes of evaluating reform programs. All these studies claim that fair chance has not been afforded compensatory and performance contracting programs. Evaluation for accountability must be improved through a more appropriate use of standardized or non-standardized instruments, better experimental designs, and more appropriate procedures for data analysis.

A unique evaluation design has been deployed in Illinois' bilingual education programs. The major thrust of this design, as the title indicates, is in instrument assessment and in varying the quasi-experimental designs. In addition to a discussion of these two areas, this report will touch on a number of factors involved in developing the evaluation design.

The importance of evaluating bilingual programs has been given very high priority. Even before the Illinois legislature passed the bills which would appropriate funds for bilingual education (the governor subsequently signed them into law in September of 1971), acknowledged authorities in evaluation design were consulted by the newly formed Bilingual Education Section of the Office of the Superintendent of Public Instruction. Among those experts who gave of their time were: Donald T. Campbell, Thomas Cook, Philip Brickman, and Lee Secrest--all from the social psychology department of Northwestern University; Marilynn B. Brewer from the psychology department of Loyola University; G. Richard Tucker and Wallace Lambert, psycholinguistists from McGill University; and Robert Cooper, a linguist from Stanford University.

Four general recommendations emerged from these consultations:

First, that prior to implementing a bilingual program in a community a sociolinguistic survey be conducted there;

second, that priority be given to early childhood programs, preferably pre-school and kindergarten;

third, that "standardized" instruments, rather than criterion-referenced tests, be selected as measurement tools;

fourth, that insofar as possible, a true experimental evaluation design be employed, with randomly assigned treatment and control groups.

This paper will discuss what was planned for the state-funded bilingual programs in each of these four areas, with most of the discussion centering on the areas of instrumentation and design. Evaluation findings are not reported in this paper.

The evaluation plans described here were developed principally in the five months in 1971 which preceded implementation of the bilingual programs; the design has been "tuned up" periodically since then. The evaluation design developed during this period was to be deployed for the first two years of the programs' existence, fiscal years 1972-73. The emphasis is heavily on a method to ascertain whether cognitive achievement is enhanced by attending a bilingual program. The important area of affective growth will be deferred to a later period of inquiry due to the scarcity of adequate attitudinal measures appropriate for Illinois "bilingual" children and to the pressing need to determine how academic achievement was affected by the program. (While supporters of bilingual programs were decidedly interested in how self-esteem is affected by the program, those who were reserving their support were much more concerned about cognitive developments.)

Sociolinguistic Surveys.

A sociolinguistic survey was not conducted prior to implementation of bilingual programs. Both advantages and drawbacks of such surveys were discussed. The advantages of conducting a sociolinguistic survey among the target communities were: (1) It could provide a means of data collection on variables whose description were important to the evaluation design; (2) it could provide information relevant to determining program content; and (3) it could provide both a vehicle for informing the bilingual community of the possi-

bilities of initiating a bilingual program and means to gain community support of the program.

The drawbacks of conducting a sociolinguistic survey included the following: (1) Growing resentment in Spanish-speaking communities to information-gathering surveys; (2) modest expectations concerning the prospect of learning something unexpected through the survey due to the likelihood that an Illinois survey would replicate antecedent surveys; (3) the timeline imposed upon the state office by circumstance would not allow time to initiate any fundamental program changes which might be suggested by any anticipated survey findings.

Alternate ways to achieve the results looked for in a sociolinguistic survey were then proposed. Collection of demographic data would be effected with the assistance of local teachers and administrators after the program got on its feet. Bilingual balance and language domain information would be gathered through student questionnaires and recordings of student speech samples. Local communities would be informed through letters from schools, visits by bilingual teachers and aides, newspaper stories, and involvement in local bilingual advisory bodies. Program change would occur whenever input seemed to warrant it. (An assessment of the success of these alternate techniques will be made in a subsequent report.)

Early Childhood Priority.

There was general agreement both among the state staff, the state advisory council, and outside consultants, that in all probability both short term and long range effectiveness of bilingual programs would be greater on younger children. The idea was to begin a program before the all-too-common deleterious effects of regular programs take their toll. Research (Hunt, 1961:

Bloom, 1964; Karnes, Hodgins, Teska, 1969) has clearly demonstrated the early years as the most educationally formidable ones. In the area of foreign languages especially, elementary school programs have repeatedly shown this to be sound. It is at this level of education that parental interest in their children's educational development is at its most intense. Opportunities to study incremental, or follow up, effects of bilingual education are, of course, greatly enhanced by beginning programs early.

On the other hand, Illinois does not have a tradition of public preschools. Mandatory attendance begins with first grade, and up to the year 1970, local school districts were not required to provide kindergarten experience for children of parents who desired it.

It was decided to concentrate most of the resources available in FY-72 on the K-3 level. (two secondary projects were funded in Chicago.) In FY-73, a number of preschool bilingual projects were funded, and most existing K-3 programs were extended to K-6. (One additional secondary program was funded in Chicago, and one dropout prevention program was funded downstate.)

Having decided, largely because of the time factor, not to attempt a sociolinguistic survey of selected Spanish-speaking communities, and after having set priorities for funding at the primary level, our attention focused on the problem of what instruments to select to measure cognitive growth of "bilingual" children.

Selection of Instruments.

Input variables. One selects instruments to test a specific population.

The population to be tested in this case consists of Illinois children of Spanish-speaking background. Yet an educational program that works well for a Cuban youngster may not be equally effective with Chicano children. The program may be more effective with children of one age than another. Achievement of the product oriented goals listed earlier are dependent on the initial (i.e. pretest) language ability in both English and Spanish. Eight different variables which help describe the student are identified in this design as input variables:

| | |
|---|---|
| (1) Grade | Pre-school through 6th grade. |
| (2) Sex | Male and female |
| (3) District | 1 through 22 |
| (4) Treatment | Bilingual, TESL and TERC (Teaching English in Regular Classroom). |
| (5) Ethnicity | Mexican, Puerto Rican, Cuban, U. S. Latin, Other Latin, and Anglo. |
| (6) Residency in U. S. | Port of entry, 1/4th of student life, 1/2 of student life, 3/4th of student life and all of student life. |
| (7) English language proficiency | 3-point scale on teacher rating, and 10-point scale on self rating. |
| (8) Spanish language proficiency | 3-point scale on teacher rating, and 10-point scale on self rating. |

Outcome variables. In spite of the current vogue for criterion-referenced tests, the lack of agreement over what a student should be able to do after a given amount of exposure to a bilingual program made it im-

practical to base a statewide evaluation on widely disparate, and often non-existent, teacher-made or criterion-referenced tests. The general areas to be tested are identified in this design as outcome variables.

The three product oriented goals of the Illinois bilingual education programs are goals 1 through 3 listed on page two of this report. Pre to post changes in the following output variables will be evaluated.

| | | |
|---|---|---|
| (1) | Pre-school grades: | Position in the development scale (i.e., year of implementation). |
| (2) | Grades K and 1: | Basic concepts in Spanish language. |
| (3) | " | Basic concepts in English language. |
| (4) | " | Basic concepts in Mathematics, measured in Spanish. |
| (5) | " | Basic concepts in Mathematics, measured in English. |
| (6) | " | Self-concept. |
| (7) | Grades 2 through 6: | English language reading. |
| (8) | " | Spanish language reading. |
| (9) | " | Mathematics, measured bilingually. |
| (10) | Grades 2 through 4: | Self-concept. |
| (11) | Grades 5 through 6: | Self-concept. |
| (12) | " | Attitude. |
| (13) | " | Study habits. |
| (14) | " | Level of aspiration. |

Since achievement in the bilingual program is to some extent a function of pretest standing and general intelligence, verbal and non-verbal intelligence at pretest time (only FY-72), and pretest scores on dependent variables are considered convariates for the evaluation.

It seemed uneconomical to consider development of new norm-referenced instruments until an adequate assessment of existing instruments was completed. Samples were requested of every standardized test whose use was reported by a bilingual project anywhere in the U.S. (Plakos, 1971). Tests were also identified through the reviews in the Mental Measurement yearbooks (Buros, 1965, 1972) and the UCLA Center for the Study of Evaluation handbooks (1970, 1971). These instruments were classified according to what they purportedly measured and their appropriateness for children on the elementary school level. Each instrument which promised to measure something relevant to the envisioned bilingual programs was studied, item by item, by a team of bilingual-bicultural psychologists. (Rafaela Elizondo Weffer, and Ana Belkind did most of this.)

A list of the instruments which were selected for use in most of the state programs operating on the elementary level is given in Table I.

It is immediately obvious that a test instrument which assumes fluency in a language which is not understood by the testee invites gross misrepresentation of the testee's cognitive skills in areas other than language. Too, the cultural--and often linguistic--inadequacy of translated tests is widely appreciated. Then again, since no standardized instrument has been normed on Illinois' multi-ethnic children of Spanish-language background, how would test scores be interpreted?

This sticky language problem is greatly compounded by the broad continuum of fluency in both English and Spanish over which Illinois' "bilingual" children are spread. For every conceivable point on the continuum there is some child in Illinois whose relative English/Spanish fluency would place him there.

The general solution to these problems was suggested by Rafaela Elizondo de Weffer and consists of alternating the language for every other item on a number of the tests. This technique has the potential of (a) reducing test anxiety and frustrations due to weakness in one of the two languages, (b) reducing time needed for testing, (c) reducing testing cost, (d) providing data on the relative dominance of each language, as well as data on the test's content. This technique also requires bilingual test administrators, thus avoiding difficulties in communication between tester and tested. Appropriate checks to evaluate the effectiveness of this alternate language technique will be applied.

The hypotheses developed to probe the strengths and weaknesses of the selected instruments include the following:

(1) The standarized tests selected for the battery are appropriate for measuring the outcomes of bilingual programs. (Appropriateness is considered in terms of item analysis, effect of random response on score, cultural loading, and set response patterns.)

(2) Oral examinations are superior to written examinations in eliciting maximum performance in bilingual populations.

(3) Appropriate coding of circles drawn to represent self in different situations constitutes a valid measure of the relative self-esteem of bilingual students in the respective situations.

(4) Data from the Dailey Language Facility Test can be validly interpreted for degree of bilingual balance and personality characteristics as well as for language facility.

(5) In grades 2 and 3, test performance is more related to language proficiency than to grade level, contrary to the classical

construct that as grade level increases proficiency (i.e. test performance) also increases.

(6) Non-verbal tests are more appropriate than verbal tests to measure the general ability of bilingual children.

(7) Alternating items between two languages within the same test is a more effective procedure to administer tests to bilingual student populations than the single language procedure.

(8) Alternating items between two languages within the same test does not affect the reliability of the test.

(9) The sequence of the two languages in testing bilingual populations by the alternate language testing procedure does not affect the performance in either language.

(10) Scores on the numerical ability subtest of the Inter-American General Ability Test is a valid index of the mathematics achievement of bilingual students.

The testing periods were set for January, 1972, May, 1972, October, 1972, January, 1973 (for downstate only), and May, 1973. The test-taking time for each student per testing period averages two and one half hours. This is generally split between two days to avoid fatigue. Testing is administered by bilingual-bicultural testers who have been inserviced in the techniques to be used with the instruments. (The initial testing period--January, 1972--was accomplished some six weeks after commencement of the bilingual programs. An important function of this delay was to reduce testee anxiety.)

(Because of this time-series design, a report of program effects would suffer a two-year delay. To get an advance indication of how the program

was going, a preliminary evaluation report was presented. This report was based on a study of the test data of first graders from eleven downstate programs. See Weffer, 1972.)

Before test data from these instruments can be interpreted in terms of the achievement of Illinois children of Hispanic background, the reliability of the instruments must be determined. To assess reliability, KR-20 and split half techniques are being applied to each of the instruments and their subtests, and correlations determined for all instruments and subtests. Data from the first testing period is being used for this purpose. The more numerous test data of the third testing period will be used to replicate the initial findings. (First testing period data will be based exclusively on downstate scores, while the third period data will include both Chicago and downstate scores.) Finally, norms based on the performance of Illinois children of Hispanic background will be established with the data from the third testing period.

Test reliability answers the question of how dependable are the test scores. That is, how much fluctuation can be expected in a given instrument. But high test reliability does not necessarily indicate that the test is testing what the testers want it to. This is a question of test validity.

Whether in fact the selected instruments measure content and skills which are central to the objectives of bilingual program as actually implemented needs to be demonstrated. Indices of the validity of these instruments will be attempted in several ways. Test scores will be correlated with teacher grades; the purported test objectives will be assessed by teachers via questionnaires as to their relevancy; a committee of teachers will evaluate the tests on the basis of an examination of the cultural and/or linguistic biases of the test items.

Evaluation Designs.

Programs are evaluated so changes can be made which will enhance their effectiveness. Since there is widespread interest in the worth of bilingual education, an evaluation design was sought which would permit broad generalizations as to treatment effect. The fundamental policy questions to be answered were: (1) Can achievement of children of Hispanic background be adequately measured by existing standardized instruments? (The previous discussion of instrumentation deals with this point.); and (2) Do children in bilingual programs learn as much or more in the routine school subjects than they would have had they stayed in the regular school program? In addition, baseline data needs to be collected on whether the effects of a bilingual program are most noticeable during the first year or so of a child's participation, or whether the effects are incremental and whether there is a critical point for beginning bilingual education.

There are two major approaches to controlling for artifacts which lead to a distorted view of bilingual program effects. One approach employs complex statistical techniques, such as path analysis. This technique, pioneered by Otis Dudley Duncan, is exemplified in the recent study by Christopher Jencks, et al, Inequality: A Reassessment of the effect of Family and Schooling in America (1972).

The other approach is the treatment-comparison group technique. In its simplest form, equivalent subjects in experimental and control conditions are pre and post tested. The differences would then become the critical points of illumination. The best contemporary exposition of this technique was done by Campbell and Stanley (1963).

The single most potent way to increase the interpretability of a comparison-group design is to assign subjects randomly to treatment (bilingual program) and control (regular school program) conditions. Random assignment makes a "true" experimental design possible, whereas the same design with "comparable" but not randomly assigned control groups Campbell calls a "quasi-experimental" design. The results from true experimental designs are, of course, much easier to unequivocally interpret than are quasi-experimental designs. The relative strength of a quasi-experimental design depends largely on how initially equivalent the treatment and comparison groups are. (The other criterion for judging the strength of a quasi-experimental design is the number of controlled threats to internal and external validity.)

We decided to aim for a true experimental design, a la Campbell and Stanley, insofar as possible. Where random assignment was not feasible, the identification of similar but not equivalent comparison groups was attempted. Since reliability and external validity are enhanced by a large sample representing schools with differing characteristics, all state-funded bilingual programs throughout the state were to be included in the overall design. (A detailed description of the strategies employed to reduce the threats to both internal and external validity for each design, and a discussion of a unique aspect of design manipulation, is being prepared as a separate report.)

The designs as they were planned and implemented--what was implemented was not always what was planned--for each of the bilingual projects which were funded in FY-72 and/or FY-73 are presented in Table II.

Rationale for multiple designs. There are three main reasons to employ multiple overlapping designs. First, local conditions differ widely and a design feasible in one school may not be physically possible or politically desirable in another school setting. For example, in one school all eligible students may be enrolled in the program, where in another, only a fraction may be so enrolled. Second, the evaluator can never be certain in field settings that what begins as a true experiment will end up that way. Because so many field exigencies work to erode or subvert carefully controlled experimental conditions, one has to be prepared with alternate quasi-experimental designs. Third, while no quasi-experimental design adequately controls for each of the nine threats to internal validity and the three treats to external validity (see Campbell and Stanley, 1963), by overlapping the design the potential to minimize the strength of rival explanations of the data is increased. A subsequent report will discuss this in much greater detail.

Random assignment. When the degree of relative need is not considered an especially relevant criterion of inclusion in the program (due perhaps to an especially large sample size), students can be randomly selected from a list of subjects which is approximately twice the size which can be accommodated ultimately in the bilingual program.

The obvious disadvantage of this in schools without twice the number of very needy students that the program can handle is that many students who badly need the program will lose their place to others of more marginal need. Schools have not reacted enthusiastically to randomly selected treatment-control groups and this model was abandoned after an abortive try.

An additional objection to having a randomly selected control group within a school is that the students selected by schools for inclusion in bilingual education programs are generally the most needy, who, because of this, cannot be compared to a group which has less need for the program when the purpose of the comparison is to demonstrate the relative efficacy of the treatment.

Random within stratum. For FY-73, a compromise true experimental design was proposed for eight Chicago schools and two downstate schools. (This design was suggested by Donald T. Campbell.) These schools were asked to categorize their students of Hispanic background who might potentially benefit from enrollment in a bilingual program into three categories the most needy, the second most needy, and lastly, students who would presumably profit from a bilingual program but for whom there is no present hope of being included, given the limited available resources. Criteria for determining need was left to each school to determine.

A typical design of this type in a school which could handle about 150 students in their bilingual program might list 50 children in the first most-needy category, 20 in the next-most-needy category, and perhaps 500 in the least-needy category. The true experiment occurs within the second category. Here, about half of the students are randomly selected for the bilingual program. Their progress is compared to that of the other half of the same category who continue in the regular school curriculum. It will be noted that external validity is made more problematic by this design since the extremes at both ends of the need continuum have been omitted.

Parallel schools/classes. Comparisons are being attempted where program schools or classes can be matched on a number of socioeconomical

variables with nearby non-program schools or classes. There are three downstate districts with bilingual programs in some but not all of the eligible schools. In Chicago, one non-program school has been identified through matching, and two schools have identified parallel classes within the program buildings.

Regression-discontinuity. This design takes advantage of situations where a sharp arbitrary cutoff of subjects who are eligible for the bilingual program becomes necessary. One such cutoff point was the result of a policy decision to limit most programs during FY-72 to grades K-3. A second cutoff point is feasible where a school ranks each student in a given grade according to need for the program, then selects the cutoff point which separates program from non-program children. In the few instances where this type of cutoff was implemented, schools were asked to priority rank twice the number of students that the program could accomodate. Five or ten numbers on each side of the "optimum" cutoff point were then identified, and the cutoff was determined randomly within this band.

The regression-discontinuity design consists mainly in (1) obtaining test data on experimental subjects by grade level, (2) obtaining test data on subjects in adjacent grade levels which are without bilingual programs, (3) extrapolating the scoring trend of the grade levels experiencing bilingual programs to non-program levels, and (4) comparing the obtained trend for non-program grade levels with the trend obtained through extrapolation.

Grade-cohort. This design takes advantage of the fact that the test data of adjacent grade levels overlap without any systematic bias, provided the school has not previously maintained the experimental program.

A fourth grade student at the end of the academic year is expected to be at the fifth grade level as far as his academic achievement is concerned. As a corollary to this statement, a fifth grade student at the beginning of the year could be considered to be at the fourth grade level as far as academic achievement is concerned. Therefore, the pretest scores of the fifth graders can be compared to the posttest scores of the fourth graders. The same logic can be applied to the other grade levels. This method of comparison is feasible for most programs initiated in both FY-72 and FY-73.

Stratified student population. In this design, different populations are compared for their contrastive interest. Native speakers of English and native speakers of Spanish, Latins in a bilingual program and Latins not in a bilingual program, Anglos in a bilingual program and Anglos not in a bilingual program, are the contrastive categories employed in this design.

Between-groups hypotheses.

In addition to instrumentation hypotheses which have already been presented, three other types of hypotheses have been developed as part of this general evaluation design--within-program hypotheses, between-groups hypotheses, and hypotheses concerning validity threats which are affected by manipulating overlapping design. These latter hypotheses will be reported later when the multiple designs approach is explicated.

The between-groups hypotheses form the major probe area along with the instrumentation hypotheses, of the first 16 months of this design. The purpose of these between-groups hypotheses is to focus clearly on how children in bilingual programs achieve when compared to similar children who are in the regular school curriculum. These hypotheses are graphically presented in Table III.

Within-program hypotheses.

After probing the question of whether students learn more in a bilingual program than they would have had they stayed in the regular school program, there is another question to ask: How much mathematics, science, social studies, language arts did they learn in the experimental program?

The best way to get answers to these questions is through criterion-referenced tests. Unfortunately, as we have already noted, these instruments are not currently available in a form suitable for bilingual programs. In an effort to press the selected norm-referenced instruments (see Table I) into double service, a number of hypotheses were developed which attempt to exploit whatever potential these instruments hold for measuring concept mastery. A list of these hypotheses follows:

(1) Eighty percent of the students in grades K and 1, at the end of each year will show a mastery of 80 percent of the concepts tested through one or more of the following instruments.

a. BOEHM test of Basic Concepts in English (grades K-1).

b. BOEHM test of Basic Concepts in Spanish (grades K-1).

c. Test of Basic Experiences in English Language (grades K-1).

d. Test of Basic Experiences in Spanish Language (grades K-1).

e. Test of Basic Experiences in Mathematics, tested through Spanish (grades K-1).

f. Test of Basic Experiences in Mathematics, tested through English (grades K-1).

(2) Assuming that a composite score on bilingually administered Test of Basic Experiences is a measure of bilingualism, 80 percent of the students in grades K and 1, at the end of the year, will show a mastery of 80 percent of the concepts tested through the

instrument. (The assumption about the composite score will be tested through appropriate analyses of correlations among a, b, c, and d above.)

(3) Assuming that a composite score on the two forms, form A - Spanish and form B - English, of the BOEHM test of Basic Concepts is a measure of bilingualism, 80 percent of the students in grades K and 1, at the end of the year will show a mastery of 80 percent of the concepts measured by the two instruments. (The assumption about the composite score will be tested through appropriate analyses of correlations among a, b, c, and d above.)

(4) A statistically significant change beyond normal growth rates in the pre to post performance of the students in grades K and 1 will be evidenced after five to nine months participation in the bilingual program, as measured by the scores on each of the following measures:

a. BOEHM test of Basic Concepts - English

b. BOEHM test of Basic Concepts - Spanish

c. Test of Basic Experiences - English language

d. Test of Basic Experiences - Spanish Language

e. Test of Basic Experiences - Mathematics, tested through English.

f. Test of Basic Experiences - Mathematics, tested through Spanish.

(5) Participating students in grades 2 through 6 when posttested through appropriate levels of the tests, will show one month's growth from pretest status for every month of participation in the program, as measured on each of the following tests:

a. English Reading (Interamerican Series)

b. Spanish Reading (Interamerican Series: Lectura)

(6) At the end of the year, 80 percent of the students in grades 2 through 6 will show a mastery of 80 percent of the concepts tested through appropriate levels of the TOBE and BESC Math Test mathematic test

(7) Change in the performance from beginning of the year to end of year of those students who at pretest rank in the lower quartile on Self/Concept/Affective Factors test will be statistically significant at the .05 level after scores are corrected for measured regression.

Process evaluations. The whole thrust of the evaluation design described in this report is product oriented, with its concern for measured cognitive achievement among Spanish-speaking children in elementary school. Yet an evaluation of the teaching process involved in helping children achieve is clearly relevant to an understanding of the effectiveness of a bilingual program.

Two process evaluations are in operation, one is a teacher self-assessment narrative done periodically to evaluate the effectiveness of his teaching strategies in meeting each of the seven state goals of bilingual education. The second process evaluation is accomplished through onsite visitations by teams of observers. Both of these process evaluations will be described at greater length and assessed in a subsequent report.

Anticipating design refinements for FY-74. The evaluation design described in this report is envisioned as a developmental method to obtain data on questions whose focus is being continually sharpened. We already perceive a need to incorporate a greater variety of evaluative instruments

into next year's design: affective measures, new or different standardized tests, criterion-referenced instruments, diagnostic measures, and instruments appropriate for the secondary school level. Due to the heavy reliance on test instruments, unobtrusive techniques need to be developed. We anticipate short-term experiments within bilingual programs to gauge the effect of various program subcomponents.

The plans for assessing the effect on the data of instrumentation and design variation are being implemented. A later paper will assess the role played by these two procedures in increasing accountability. The question is not which design or what instrument is best for assessing bilingual education programs, but what combination of designs and what combination of instruments give the most accurate picture.

REFERENCES

Bloom,B.S. Stability and Change in Human Characteristics. New York: John Wiley and Sons, 1964.

Buros,O.K. The Seventh Mental Measurement Yearbook (Vols. I and II). New Jersey: The Gryphon Press, 1972.

Campbell,D.T., and Erlebacher,A. How regression artifacts in quasi-experimental evaluations can mistakenly make compensatory education look harmful. In J. Hellmuth,(ed), Compensatory Education: A national debate, Vol.III, Disadvantaged Child. New York: Brunner/Mazel,1970.

Campbell,D.T., and Frey,P.N. The implications of learning theory for the fade out of gains from compensatory education. In J. Hellmuth,(ed), Compensatory Education: A national debate, Vol.III, Disadvantaged Child. New York: Brunner/Mazel, 1970.

Campbell,D.T., and Stanley,J.C. Experimental and Quasi-experimental Designs in Educational Research, Chicago: Rand McNally, 1963

Fitzgibbon,T.J. The use of standardized instruments with urban and minority group pupils. Test Department, Harcourt Brace Javanovich Inc. n.d.

Garfinkel,I., and Gramlich,E.M. Astatistical analysis of the OEO experiment in educational performance contracting, OEO pamphlet 3400-6, June 1972.

Hunt,Mc. V. Intelligence and Experience. New York: Roland Press, 1961.

Karnes,M.B., Hodgins,A.S., and Teska,J.A. Investigations of classroom and at home intervention. Vol.I, Research and development on preschool disadvantaged children. Final Report. Bethesda,Md.: ERIC document reproduction, (ED036-663), 1969.

O'Connor,E.L.,and Klein,S. A statistical analysis of the OEO experiment in performance contracting. Paper presented at AERA annual convention, New Orleans, 1973.

Plakos,J. Tests in use in Title VII bilingual education programs. Fortworth, Texas: National Consortia for Bilingual Education, 1971.

UCLA-CSE. Elementary school test evaluations. Los Angeles: Center for Study of Evaluation, 1970.

UCLA-CSE. Preschool/Kindergarten test evaluations. Los Angeles: Center for Study of Evaluation, 1971.

Weffer, R.D.C.E. Effects of first language instruction in academic and psychological devdelopment of bilingual children. Doctoral dissertation, Illinois Institute of Technology, 1972

Wrightstone, J.W., Hogan,T.P., and Abbot,M.M. Accountability and associated measurement problems. n.d. Test Department, Harcourt Brace Javanovich Inc.

TABLE I – DESCRIPTION OF INSTRUMENTS

| Measuring Instrument | Language of Instrument | Level | Grade | 1/72 I | 5/72 II |
|---|---|---|---|---|---|
| Test of Basic Experiences-Language | Eng/Span | K | Kinder | X | X |
| Test of Basic Experiences-Language | Eng/Span | L | 1-2 | X | X |
| Test of Basic Experiences-Mathematics | Eng/Span | K | Kinder | | |
| Test of Basic Experiences-Mathematics | Eng/Span | L | 1-2 | X | X |
| BOEHM Test of Basic Concepts Form A | Spanish | – | K-2 | | |
| BOEHM Test of Basic Concepts Form B | English | – | K-2 | | |
| Inter-American - Test of Reading | English | 1 | 1 | | |
| Inter-American - Test of Reading | English | 2 | 2-3 | X | X |
| Inter-American - Test of Reading | English | 3 | 4-5-6 | X | X |
| Inter-American - Test of Reading | English | 4 | 7-8 | | |
| Inter-American - Prueba de Lectura | Spanish | 1 | 1 | | |
| Inter-American - Prueba de Lectura | Spanish | 2 | 2-3 | X | X |
| Inter-American - Prueba de Lectura | Spanish | 3 | 4-5-6 | X | X |
| Inter-American - Prueba de Lectura | Spanish | 4 | 7-8 | | |
| Inter-American - General Ability | Eng/Span | 1 | 1 | | |
| Inter-American - General Ability | Eng/Span | 2 | 2-3 | X | X |
| Inter-American - General Ability | Eng/Span | 3 | 4-5-6 | X | X |
| Inter-American - General Ability | Eng/Span | 4 | 7-8 | | |
| Dailey Lang. Facility Test | Eng/Span | – | K-1 | X | X |
| BESC - Draw-a-Circle Self-Concept | Eng/Span | – | K-3 | X | X |
| BESC - Language Usage Questionnaire | Eng/Span | – | K-3 | X | |
| BESC - Demographic Questionnaire | Eng/Span | – | K-6 | | |
| Chicago Self-Concept Scale | Eng/Span | – | K-4 | | |
| BESC - Test of Basic Mathematics | Eng/Span | 1 | 2-3 | | |
| BESC - Test of Basic Mathematics | Eng/Span | 2 | 4-6 | | |
| BESC - Test of Basic Mathematics | Eng/Span | 3 | 7-8 | | |

TABLE I – DESCRIPTION OF INSTRUMENTS

| | Language of Instrument | Level | Grade | Testing Period | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | | 1/72 I | 5/72 II | 9/72 III | 1/73 IV | 5/73 V |
| ge | Eng/Span | K | Kinder | X | X | X | X | X |
| ge | Eng/Span | L | 1-2 | X | X | X | X | X |
| matics | Eng/Span | K | Kinder | | | X | X | X |
| matics | Eng/Span | L | 1-2 | X | X | X | X | X |
| Form A | Spanish | – | K-2 | | | X | X | X |
| Form B | English | – | K-2 | | | X | X | X |
| | English | 1 | 1 | | | | | |
| | English | 2 | 2-3 | X | X | X | X | X |
| | English | 3 | 4-5-6 | X | X | X | X | X |
| | English | 4 | 7-8 | | | X | X | X |
| ra | Spanish | 1 | 1 | | | | | |
| ra | Spanish | 2 | 2-3 | X | X | X | X | X |
| ra | Spanish | 3 | 4-5-6 | X | X | X | X | X |
| ra | Spanish | 4 | 7-8 | | | X | X | X |
| | Eng/Span | 1 | 1 | | | | | |
| | Eng/Span | 2 | 2-3 | X | X | | | |
| | Eng/Span | 3 | 4-5-6 | X | X | | | |
| | Eng/Span | 4 | 7-8 | | | | | |
| | Eng/Span | – | K-1 | X | X | | | |
| ot | Eng/Span | – | K-3 | X | X | | | |
| naire | Eng/Span | – | K-3 | X | | | | |
| re | Eng/Span | – | K-6 | | | | | X |
| | Eng/Span | – | K-4 | | | | X | |
| | Eng/Span | 1 | 2-3 | | | X | X | X |
| | Eng/Span | 2 | 4-6 | | | X | X | X |
| | Eng/Span | 3 | 7-8 | | | X | X | X |

TABLE II
STATEWIDE EVALUATION DESIGNS
AND PROJECT SITES

| Type of Comparison | FY 72 | | FY 73 | | FY 74 | |
|---|---|---|---|---|---|---|
| | Downstate | Chicago | Downstate | Chicago | Downstate | Chicago |
| I Random Assignment | 1 | 2 | | | | |
| II Random within Stratum | | | 3 | 4 | | |
| III Parallel Schools or Classes | 5 | 6 | 7 | 8 | | |
| IV Regression Discontinuity<br>A. Program, Nonprogram Grades | | | 9 | | | |
| B. Random Cutoff on Needs Scale | | | | 10 | | |
| V Grade Cohort | 11 | 12 | 13 | 14 | | |
| VI Stratified Student Population | | | 15 | 16 | | |

1. Bensenville.
2. Bowen, Burns, Cooper Upper, Sheridan, and Sullivan.
3. Bensenville.
4. Agassiz, Bowen, Burns, Cooper Lower, Gary, Komensky, McCormick, Sullivan, and Thorp.
5. Elgin, Joliet, Steger, and Waukegan.
6. Agassiz, Bowen, Burns, Cooper Primary, Cooper Upper, Lakeview, Nash, Sheridan, Sullivan, and Headley-C.
7. Joliet (Keith-C, Lincoln, Marsh-C, Marshall-C, and Parks).
8. Lowell and Sheridan.
9. Aurora, Bensenville, Chicago Heights, Des Plaines, Dundee, Elgin, Joliet, Moline, Steger, Waukegan, and West Chicago.
10. Irving and Nettlehorst.
11. Aurora, Bensenville, Chicago Heights, Des Plaines, Dundee, Elgin, Joliet, Moline, Steger, Waukegan, and West Chicago.
12. Agassiz, Bowen, Burns, Cooper Primary, Cooper Upper, Lakeview, Nash, Sheridan, and Sullivan.
13. Arcola, Crete-Monee, Danville, Elkgrove, Marengo, Maywood, Palatine, Rockford, and Wheeling.
14. Gary, Hamline, Irving, Jungman, Komensky, Lemoyne, McCormick, Morris, Nettlehorst, Plamandon, and Thorp.
15. Elgin, Joliet, Waukegan, West Chicago, Danville, Elkgrove, Crete-Monee, and Rockford.
16. In program Latins, Not in program Latins, In program Anglos, and Not in program Anglos. (Sample from Chicago Public Schools student population in program area.)

* C = Comparison School.

## TABLE III – BETWEEN-GROUPS HYPOTHESES

| Comparison | Expected Result | | | |
|---|---|---|---|---|
| | English | Spanish | Mathematics | Self-Concept |
| 1. Change in performance of students in Experimental II, compared to that of students in comparison II between two testing sessions will be . . . | Not Different | Superior | Superior | Superior |
| 2. Change in performance of in-program Latins in comparison to that of not in-program Latins between two testing sessions will be . . . | Superior | Superior | Superior | Superior |
| 3. Change in performance on in-program Latins in comparison to that of in-program Anglos between two testing sessions will be . . . | Not Different | Not Different | Not Different | Not Different |
| 4. Change in performance of in-program Latins in comparison to that of not in-program Anglos between two testing sessions will be . . . | Not Different | Superior | Not Different | Not Different |
| 5. Change in performance of bilingual students in comparison to that of Latin students in TESL Programs between two testing sessions will be . . . | Not Different | Superior | Superior | Superior |
| 6. Change in performance of bilingual students in comparison to that of Latin students in regular English classrooms, between two testings, will be . . . | Superior | Superior | Superior | Superior |
| 7. Change in performance of the experimentals group in the period between 3rd and 4th testing, in comparison to that in the period between 4th and 5th testing will be . . . | Not Different | Not Different | Not Different | Not Different |
| 8. Change in performance of in-program grade cohort in comparison to that of corresponding not in-program grades cohort will be . . . | Superior | Superior | Superior | Superior |
| 9. Change in interpolated in-program performance of 4-5-6 graders in comparison to that in observed performance of not in-program 4-5-6 graders; will be . . . | Superior | Superior | Superior | Superior |
| 10. Change in performance on in-program K-3 graders in comparison to that in interpolated not in-programs performance of K-3 graders, between two testing sessions will be . . . | Superior | Superior | Superior | Superior |
| 11. Change in interpolated in-program performance of the students (in the upper half of the need scale), in comparison to that of the observed performance of the not in-program students (in the lower half of the need scale) will be . . . | Superior | Superior | Superior | Superior |
| 12. Change in observed performance of the in-program students (in the upper half of the need scale), in comparison to that of the interpolated not in-program performance of the students (in the lower half of the need scale will be . . . | Superior | Superior | Superior | Superior |