DOCUMENT RESUME

ED 108 396 EC 072 854

AUTHOR Proger, Barton B.
TITLE A Formal Program Evaluation Model for the Special Education Programs in Pennsylvania.
INSTITUTION Pennsylvania State Dept. of Education, Harrisburg. Bureau of Special and Compensatory Education.
SPONS AGENCY Office of Education (DHEW), Washington, D.C.
PUB DATE 71
NOTE 113p.

EDRS PRICE MF-$0.76 HC-$5.70 PLUS POSTAGE
DESCRIPTORS Criterion Referenced Tests; Exceptional Child Education; *Handicapped Children; *Models; Norm Referenced Tests; *Program Evaluation; *Special Education; State Programs
IDENTIFIERS *Pennsylvania

ABSTRACT

The preliminary draft of a formal program evaluation model for special education operations in Pennsylvania is presented. Beginning chapters provide a review of literature on norm-referenced measurement, an illustration of formal and informal program evaluation in a learning disabilities context, descriptions of general implementation strategies of norm-referenced measurement, and analyses of the relationships of norm-referenced measurement to existing statewide and national assessment schemes. Subsequent chapters include a review of literature on the use of criterion-referenced measurement in formal program evaluation, a description of a criterion-referenced measurement system said to be suitable for special state-connected projects such as the National Regional Resources Center of Pennsylvania, and an explanation of machinery needed for implementation of the formal program evaluation system for special education at the state level (personnel and data banking activities). Appendixes suggest priorities in the dissemination of the draft document, provide guidelines for professional usage of accountability data at local or state levels with either total program evaluation or individual achievement monitoring, analyze possible interrelationships among existing agencies in carrying out a statewide formal program evaluation system, and outline operational steps needed to implement a statewide formal program evaluation system in its first year. (GW)

A FORMAL PROGRAM EVALUATION MODEL FOR THE SPECIAL EDUCATION PROGRAMS IN PENNSYLVANIA

A FORMAL PROGRAM EVALUATION MODEL FOR THE SPECIAL EDUCATION PROGRAMS IN PENNSYLVANIA

---:---

---:---

by

Barton B. Proger, Ed. D.
Director of Evaluation and Dissemination
Pennsylvania Resources and Information
Center for Special Education
(PRISE)

---:---

Commissioned by Dr. William F. Ohrtman, Director, Bureau of Special Education, Commonwealth of Pennsylvania, Harrisburg, Pennsylvania, and completed without special funding as a regular project activity of PRISE, 443 South Gulph Road, King of Prussia, Pennsylvania 19406. However, no endorsement of the views contained herein is to be inferred on the parts of Dr. Ohrtman, the Pennsylvania Bureau of Special Education, PRISE, or the latter's funding agency, the United States Office of Education. The views expressed are solely those of the author himself.

---:---

NOTE

The construction of this model by PRISE represents one of many separate but coordinated efforts undertaken by Dr. Ohrtman, Director of the Bureau of Special Education, Commonwealth of Pennsylvania, Harrisburg. Another evaluation-oriented activity that concerns special education in Pennsylvania is the series of discussions held by the Subcommittee on Evaluation, chaired by Dr. Richard K. Meyers, Special Education Department, Slippery Rock State College; this Subcommittee in turn is part of the State Advisory Board for Special Education.

TABLE OF CONTENTS

PREFACE

The preliminary draft of a formal program evaluation model for special education operations in the Commonwealth of Pennsylvania is presented herein. As stated in the introduction, it should be remembered that this model is only a tentative one, quite open to change. The major purpose of this document is to provide a foundation for discussion. Many changes are anticipated. One should take particular notice of the fact that this model deals with only formal evaluation of programs in terms of commonly recognized measuring instruments; at no point was subjective accreditation-type program evaluation brought into the model.

Finally, a few words are in order regarding how this model was brought into being. The model has been constructed without special funds of any type under the regular auspices of PRISE, the Pennsylvania Resources and Information Center for Special Education. PRISE is funded under Title III of the Elementary and Secondary Education Act of 1965, and is located in King of Prussia with RRC, the Regional Resources Center of Eastery Pennsylvania for Special Education. Part of the regular functions of PRISE include serving directly the members of the Bureau of Special Education in Harrisburg. Thus, Dr. William F. Ohrtman, Director of the Bureau of Special Education, came to PRISE around February, 1970, and asked that its personnel begin putting together a formal program evaluation model for consideration by the Bureau. This task was assigned to Dr. Barton B. Proger, Director of Evaluation and Dissemination for PRISE. It should also be realized that this model had to be constructed amidst the several other regular activities of RRC and PRISE. Thus, the previous summer (1970) and the present academic year (1970-1971) had been allotted as sufficient working time to provide the model.

Robert L. Kalapos
Director, RRC

INTRODUCTION : QUALIFICATIONS OF THE MODEL AND ANTICIPATION OF CRITICISMS

Dr. William F. Ohrtman, Director of the Bureau of Special Education, has been long interested in formal evaluation of programs. It is important at the outset to distinguish between informal, subjective, accreditation-type program evaluation (which is currently being carried out by the Bureau) and the more formalized, objective, program evaluation that relies on commonly recognized measuring devices. Further, individual pupil, psychological evaluations of children are not to be confused with either accreditation program evaluation or formal program evaluation. Because the implementation of formal program evaluation has been almost totally lacking in state special education operations across the nation, little help exists in the form of guidelines which might aid the Bureau in considering what the relative advantages and disadvantages of formal program evaluation are. Thus, Dr. Ohrtman assigned me (a) to gather together a state-of-the-art paper on formal program evaluation, and (b) to develop a model for formal program evaluation that the state might consider.

It should be established immediately in the minds of the readers that I have a bias toward formal measurement procedures. I believe the single most important indicant of the success of any special education program is how well children have progressed over specified periods of time in highly specific areas of behavior. To me, any program evaluation made in terms of one point in time (such as the once-a-year, standardized testing programs given in spring in regular education) offers very little in data for judging the effectiveness of programs. The once-a-year testing merely tells school personnel where the child is, with no indication of how school itself

affected those test results. One must at least have baselines against which to measure progress. Thus, I have endeavored to embody the best measurement methodology in the model developed here.

The state-of-the-art paper on formal program evaluation is derived in part from an article on accountability that was requested of me by the Journal of Learning Disabilities. The other component of the evaluation package presented here, the individual achievement monitoring system, has been devised by Dr. Lester Mann and I for the Eastern Suburban Division of the National Regional Resources Center of Pennsylvania. The monitoring system has received publicity at the Council for Exceptional Children convention in Miami in 1971 (Mann and Proger, 1971) and in the Journal of Special Education (Mann, in press). The individual achievement monitoring system is mentioned in the context of this model because it has many implications for any formal program evaluation system. The monitoring system described here embodies a great many new measurement concepts taken from criterion-referenced measurement theory.

The whole notion of formal program evaluation is related to the concept of "accountability." Unfortunately, accountability has been viewed with a type of funnel vision as being associated only with guaranteed performance contracts between publishers of educational materials and school systems. Such contracts assume various forms; some merely provide materials only, with no insurance or guarantee that certain minimum achievement will be produced in the children who use them; others provide a far-ranging package of not only materials, but also personnel, consultation, guarantees, etc. However, no matter what arrangement is reached between the publisher and the shcool system, there is an implied or stated assumption that "accountability" will prevail. Stated very simply, the term refers to the

attempt to evaluate how well certain educational objectives were achieved. Another term which is receiving increased usage in the educational literature is "program evaluation"; indeed, one might consider this term synonymous with accountability.

The sad part of viewing accountability or formal program evaluation in connection with only performance contracting, is that the vast majority of routinely funded, locally run special education programs go without any formal evaluation. The view taken in this document is that formal program evaluation should extend to all types of special education programs.

As one reads through this model, he will see that a certain amount of research design has been thrown into the total picture. Obviously, any data obtained in realistic, on-going special education programs will not be as methodologically "clean" as desired from a research point of view. There will always be the criticisms from skeptics that feel contaminated data should not be used at all; it is precisely this negative view that has kept formal program evaluation from ever reaching fruition. Nonetheless, whenever a formal program evaluation system is attempted for special education, officials must realize that such criticisms will continuously be made.

Besides the formal program evaluation system described herein and the individual achievement monitoring system, the reader will see in the appendix to this report that part of the "machinery" needed at the state level to make such evaluation systems work is a data-banking activity. A large data processing and data storage arm is needed; these facilities are usually already available at local and state levels. Arrangements would have to be made with existing staff to work out a base of cooperation. In connection with such data-banking activities, there will no doubt be cries of "invasion of privacy" and "everything our children has ever done has been reduced to a mass of numbers on computer cards.". True, such dangers

are always present, but it is also felt that when and if the Bureau of Special Education finally feels it has the necessary sophistication in the model so that it would want to recommend it, then safeguards on interpretation will have to be implemented.

With regard to interpretation and use of accountability or program evaluation data, gross misconceptions about accountability in general have given rise to unwarranted criticisms of systems for program evaluation systems. Accountability has been twisted to mean that unflattering results in any given program might even cause a teacher or administrator who has been associated with that program to be reprimanded, to lose pay, or even to be fired. This view could not be farther from the truth. If I were interpreting data obtained throughout the Commonwealth from on-going special education programs, first, at the local level, no teacher within a school organization would be compared favorably or unfavorably with any other teacher. Such comparisons are NOT the purpose of formal program evaluation, although many misguided people have attempted to convey this threatening image. As I see it, the major goal of accountability is to look at the overall progress of children within one major programming approach (if only one approach is used) or to compare one programming technique with one or more others. (if more than one approach is used with the same children). Second, at the state level of data-banking, a school's (or I.U.'s) program for, say, the trainable, in one part of the state will NOT be compared favorably or unfavorably with a similar program in another part of the state. This is not meant to be a vehicle for approving or disapproving on-going programs, for supplementing or detracting from federal aid to such programs. It must be remembered that the number of confounding variables in making such threatening comparisons is far too great to allow valid comparisons of that type. The main point is that only programming techniques as such (not an individual

teacher, administrator, or school organization) are on trial or held "accountable." It is the hope of the Bureau of Special Education that feedback will be gotten from the data-banking activities for making decisions on w[illegible]o keep an on-going programming technique or to change to a different one. Personnel -- administrators, teachers, etc. -- should not feel threatened in the least.

The basic philosophy of holding individual educational staff members accountable for their action or lack of action does, of course, have merit. However, the truth must be faced that accountability machinery is just not yet that refined for making such finely honed decisions. Nonetheless, many benefits for decision-making can be gotten from the existing potential in formal program evaluation. Such benefits are explained in the enclosed model in great detail. In brief, however, data-banking activities for accountability are meant as an initial effort at the state level to provide answers to questions such as: (a) How far can children within given ranges of potential and with specified disabilities progress over a certain amount of time? (b) How much farther or less can such a child progress under a different instructional approach? (c) What cost-effectiveness factors enter the picture? The answer to (a), as simple a question as it is, is unknown for any area of exceptionality. Permanent records will be kept on the answers to these questions, as well as others. As the data accumulates to a greater extent, more complex questions can be answered. This is the type of monitoring job a state can be doing if it so desires.

In this somewhat lengthy introduction, I have endeavored to give the reader a flavor of the underlying philosophies that guided the development of the enclosed formal program evaluation model. The implementation of a decently functioning accountability system is a vast undertaking. With such a huge job, the whole range of measurement criticisms will be met. I just

hope that special educators will not be distracted by potential criticisms to such an extent that they fail to see the forest for the trees in terms of long-range benefits.

Barton B. Proger, Ed.D.
Director of Evaluation
and Dissemination, PRISE

CHAPTER I

FORMAL PROGRAM EVALUATION (NORM-REFERENCED MEASUREMENT): REVIEW OF LITERATURE

Introduction

In this first chapter, several types of evaluation procedures will be surveyed and put into perspective with respect to one another. There are at least five major types of evaluation activities that are said -- rightly or not -- to fall within the province of formal program evaluation: (a) formal program evaluation (norm-referenced measurement), (b) individual achievement monitoring (criterion-referenced measurement), (c) accreditation-type on-site evaluation visits, (d) descriptive systems-analyses evaluations, and (c) demographic data record keeping. In turn, these five evaluation activities can be envisioned to occur at four levels: (a) national, (b) state, (c) regional, and (d) local. These preliminary relationships are indicated in the matrix in Figure 1.

The present chapter will focus on only the first type of evaluation system: formal program evaluation (norm-referenced measurement). However, passing mention in this chapter will be accorded to accreditation on-site visits, descriptive systems analyses, and demographic record keeping; these three types of evaluation systems do have some limited value and in some situations may even be deemed necessary. Nonetheless, it is the opinion of the author that only formal program evaluation and individual achievement monitoring really deserve any sustained attention when educational agencies are considering the implementation of sophisticated and worthwhile evaluation systems. Thus, a separate chapter will be devoted later to ex-

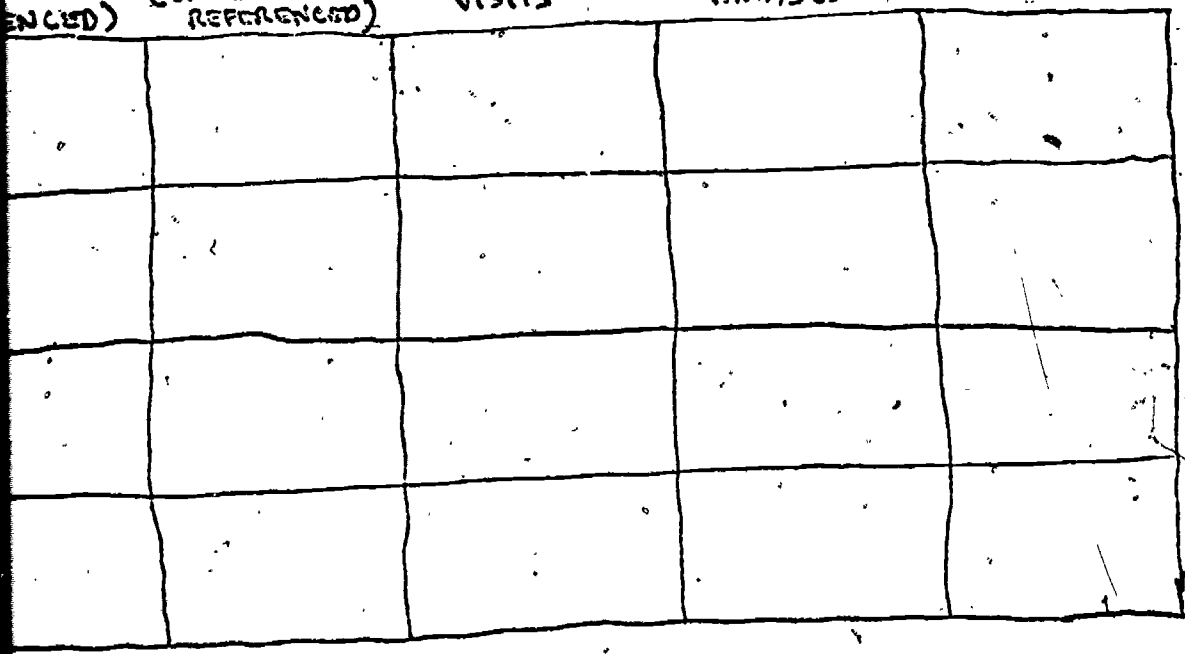
FIGURE 1

OUTLINE OF MAJOR TYPES OF EVALUATION SYST

	FORMAL PROGRAM EVALUATION (NORM-REFERENCED)	INDIVIDUAL ACHIEVEMENT MONITORING (CRITERION-REFERENCED)	ACCREDITATION ON-SITE VISITS	DESCRIPTIVE SYSTEMS ANAYSIS	DEMOGR RECORD KEEPIN
NATIONAL					
STATE					
REGIONAL					
LOCAL					

FIGURE 1

LINE OF MAJOR TYPES OF EVALUATION SYSTEMS



plaining the merits of individual achievement monitoring.

This paper will consider the issue of program evaluation in the area of learning disabilities. Several topics will be covered: (a) definitions of program evaluation, (b) currently used models of program evaluation, (c) a critique of those models, (d) an example of both formal and informal program evaluation in the field of learning disabilities, (e) a suggested resolution of the program evaluation dilemma, and (f) some hints of the future in learning disabilities program evaluation.

DEFINITIONS AND NATURE OF THE PROBLEM

In the field of learning disabilities, there is a great deal of confusion about what "evaluation" means. Indeed, this confusion extends into all areas of exceptionality and even into regular education. Part of the confusion stems from the great deal of emphasis given to clinical evaluations or diagnoses of individual children. Too often special educators have considered individual pupil evaluation to be synonymous with program evaluation.

Before going any further, a working definition of program evaluation must be given. I consider program evaluation to be the process of gathering evidence (test data, anecdotal teacher records, clinician observations, and so on) on the effectiveness of the total learning disabilities program (wheather run by an individual public school district, a private or parochial school, a county school system or an intermediate unit, or even a state hospital). To guage the effectiveness of the program as a whole, program evaluation relies on the individual pupil diagnoses or evaluations. This information on individual pupils is combined or averaged in meaningful ways to gauge the progress of certain types of pupils within the learning disabled program. Such pooled information will yield results in a more manageable form than the separate pupil evaluation records so that the program administrator, teacher, or other staff member can make future programming decisions on a rational basis. Too

often educators have been accused of making major programming decisions on an intuitive basis -- "armchair philosophizing" (cf. Proger et al., 1970). Program evaluation, if used intelligently, can help eliminate such criticism. In summary, program evaluation can be considered a step above pupil evaluation in complexity.

To clarify further the working definition of program evaluation, one need only look at the federally funded programs under the Elementary and Secondary Education Act of 1965. Projects funded under Title III of that act are required to submit formal program evaluation data on the effectiveness of the program with children. Program evaluation has received increasing attention recently as federal and state officials become more and more aware of the low quality -- or even complete absence -- of program evaluation in federally supported projects (Smith and Brecknell, 1969; Erickson, 1970). Thus, certain sources of federal and even state funding have required learning disabilities educators to produce at least some semblance of program evaluation, no matter how poor or inappropriate. However, let us not delude ourselves. It is in the area of locally funded learning disabilities programs that program evaluation is most crucially needed and, paradoxically, most frequently absent!

Program evaluation is also frequently confused with accreditation of schools (e.g., North Central Association of Colleges and Secondary Schools, 1969). While the opinions of visiting experts, teachers, students, and parents are important, accreditation forms a distinct area of the broad field of evaluation that will not be considered in this paper. Further, the discussion of program evaluation here will be confined to student achievement, feelings, and performance. Teacher competencies, financial resources, organizational structure, etc., are left to other types of evaluation experts. I believe pupil functioning is the single most important aspect of any learning

disabilities program to be examined. Program administrators and teachers will be able to obtain a great deal more detailed information for decision making from direct data on the pupils as compared to a model weighted down with other variables such as money and staff competencies. The completely generalized, competently functioning program evaluation model is far in the future, to say the least!

The reader should also be aware that several program evaluators (Scriven, 1967; Stake, 1967, pp.525-526; Atkinson, 1967, p.2) suggest that the process of program evaluation should not only describe the change that occurs in pupils over the course of time (such as gains in test scores) but should also judge whether those changes are acceptable or not. Some might question this viewpoint in that it infringes upon the nonevaluator-program administrator's role of decision making. Other definitions of program evaluation have been posited (Cronbach, 1963, p.672; Griessman, 1969, p.17; Welch, 1969, p.429). However, the stage is now set for examining some of the major program evaluation models.

CURRENTLY USED PROGRAM EVALUATION MODELS

How does one go about designing an adequate program evaluation scheme? There are many models now available that describe the major steps in program evaluation design; as generalized guidelines, these descriptions deserve the name of "models." One should not expect to see a complicated mathematical model; such models have had success only in very specific contexts that do not possess the many complexities of an ongoing learning disabilities program (cf. Welty, 1969; Brooks, 1969; Alkin, Glinski, and Wininger, 1969).

Perhaps the most influential program evaluation model to date has been the Stufflebeam (1967) CIPP structure; Context Input, Process, and Product. Rather than review some of the characteristics of the original CIPP model, let us look at one of the many, modern adaptations of the model: the CDPP

represents Context, Design, Process, and Product. Describing the major steps in the CDPP evaluation process, Randall (1969, pp. 40-42) states: "Context evaluation consists of planning decisions and context information that serves them ... Design evaluation entails structuring decisions which depend on design information ... the objectives need to be specified operationally if possible, and activities or means of attaining them need to be specified ... After a design has been structured and is put on trial, often called the pilot test, restructuring decisions are faced. Restructuring decisions are based on process information ... After components of a design have been tested, they can be put together in a program for a product or field test. Since this is the first full-cycle test, the major decisions faced are whether to recycle through another full-scale field test. The information needed, called product information, entails not only evidence about effectiveness in attaining short - and long-range goals, but also effectiveness ... compared with that of another program or strategy."

Because the CIPP-CDPP-type model is the basis for many other program evaluation schemes, it would be helpful to illustrate the CDPP variation briefly with a learning disabilities problem. Let us suppose a perceptual motor training program has been deemed appropriate for use with a certain group of learning disabled children in a private school. The Context part of the evaluation cycle would involve looking at the needs of the children in detail and the associated problems behind those needs. Preliminary studies (such as individual pupil diagnoses) would be appropriate to help determine needs of the pupils. On the bases of the pupil needs and the problems underlying them, broad program goals and specific behavioral objectives are determined. Design evaluation (or in Stufflebeam's terms, Input evaluation) can be thought of as having a primary purpose of arriving at a feasible training program for the learning disabled pupils. Design evaluation entails con-

sideration of the constraints of the private school: funds, staff skills, facilities, scheduling, etc. The program administrators and other staff members have the responsibility for examining all the specific details of the perceptual training program they originally thought appropriate; the literature -- both research and philosophical opinion -- would be searched thoroughly to gain insight into the virtues and flaws of that particular training program. At the same time, however, alternative training programs would be examined in the literature to see if an even more suitable program might be found (see Proger et al., 1970). Once these preliminary steps have suggested which perceptual training program might be most suitable with the children under the constraints of the private school, design evaluation concludes its role by specifying more fully what is to be done with the children. That is, design evaluation also implies the specification of the actual steps to be used in the training program finally selected, and the specification of a design for gathering evidence of effectiveness. Process evaluation can refer to an actual pilot test of the perceptual motor training program; evidence of effectiveness during the pilot test is used to restructure the final program that is to be used later in the regular activities of the private school. However, more often than not, pilot testing will not be possible and process evaluation will refer to in-process quality control monitoring of the final program itself; evidence will be gathered systematically throughout the actual training program and will be used to restructure the program as it is running. Product evaluation is perhaps the most familiar step, since it refers to gathering the evidence of effectiveness at the end of the perceptual motor training program. Thus, when product data is compared to pretest data and to process data, analyses can be generated which yield information that the program administrator can use as a basis for making future decisions.

Thus, one sees that the CDPP program evaluation model generally fits all aspects of the evaluation process of any learning disabilities program. This is one of the primary strengths of the CDPP model. A more detailed discussion of context evaluation can be found in Freedman and Swanson (1969) and in Hammond (1969). Stufflebeam (1969) provides a recent discussion of his CIPP model.

Another well-known program evaluation model is that of EPIC (Evaluative Programs for Innovative Curriculums). The EPIC "structure for evaluation" (Hammond, n.d.; EPIC, 1968) consists of a three-dimensional figure: a cube. The "Behavior" edge is divided into three units: cognitive domain, affective domain, and psychomotor domain. The "Instruction" edge has five units: organization, content, method, facilities, and cost. The third edge is labeled "Institution" and is divided into six units: student, teacher, administrator, educational specialist, family, and community. The cube is embedded into a five-category scheme of variables. The first category -- "Prediction Sources" -- implies that one examines the various types of instruction that might be used in a given situation. The second category of "Descriptive Variables" suggests that the actual steps to be used in the instructional techniques are to be specified carefully, along with the constraints that the institution places upon the teaching. "Objectives" forms the third category of variables. The fourth category consists of the cube described above. The actual design for collecting the effectiveness data is specified in this step. The fifth category -- "Criteria of Effectiveness" -- implies the analysis of all data obtained. One can see the similarities between the CDPP model and the EPIC scheme.

The reader is now aware of two main program evaluation models and the types of guidelines suggested by each. Schematically, the EPIC design can be said to be representative of the geometric model-building efforts (in this

case, a cube), while the CIPP design characterizes the logical eggcrate pattern (the four main stages are placed horizontally on top of a rectangle, and subdivisions are placed vertically down the left side of the rectangle, thus forming an eggcrate classification scheme). However, a few other models might be mentioned here for further reference.

Scriven (1967) has produced a lengthy book chapter on what he envisions as program evaluation. He tries to formalize in much greater detail than other writers in the evaluation field what he considers to be the "methodology" of evaluation. One gets into statistical design discussions and other technical areas.

Stake (1967) builds a logical-eggcrate design. His basic model consists of two major blocks of information: a "Description Matrix" and a "Judgment Matrix." Data can be subclassified in either matrix as "antecedents", "transactions", or "outcomes." The descriptive matrix is further subclassified into "Intents" and "Observations", while the corresponding dimension in the judgment matrix consists of "Standards" and "Judgments."

Atkinson (1967) divides his evaluation model into three domains according to the areas objectives can be constructed: structure (school plant, organization,, etc.), process (instruction), and product (student outcomes in behavior).

Pohland (1970) describes a geometrical program evaluation model developed by Howard Russell and Louis Smith at the Central Midwestern Regional Educational Laboratory, Inc., St. Ann, Missouri. Like the EPIC design, the CEMREL model is a three-dimensional cube. Along one edge of the cube is the "focus" of evaluation (student, mediator, or material). The second edge deals with the "role" of evaluation (formative and summative). The final edge consists of "data" (scale measures, questionnaire responses, and participant observations).

A discussion of basic program evaluation models would be incomplete without systems analysis. In 1968 a group commissioned by the National Security Industrial Association studied the application of systems analysis in defense to the area of education. Carter (1969, pp. 22-23) summarized eight steps of systems analysis that could be useful in education: "(a) State the real NEED you are trying to satisfy; (b) Define the educational OBJECTIVES which will contribute to satisfying the real need; (c) Define those real-world limiting CONSTRAINTS which any proposed system must satisfy; (d) Generate many different ALTERNATIVE systems; (e) SELECT the best alternative(s) by careful analysis; (f) IMPLEMENT the selected alternative(s) for testing; (g) Perform a thorough EVALUATION of the experimental system; (h) Based on experimental and real world results, FEEDBACK the required MODIFICATIONS and continue this cycle until the objectives have been attained." Robertson (1969, pp.31) claims " ... application of systems analysis techniques to evaluation differs from PERT (Program Evaluation Review Technique). PERT focuses on the steps, the time, and other expenditures in the identified evaluation or research processes, while ...[systems analysis] should be thought of in terms of the operating program, not the evaluation process per se." Additional thoughts on systems analysis models can be found in Dyer (1969); Ammentorp, Daley, and Evans (1969); Ryan (1969); Wallace and Shavelson (1970).

CRITIQUE OF PROGRAM EVALUATION MODELS

By now I hope to have conveyed to the reader the trend in current educational literature concerning the construction of general program evaluation models. In recent years educators of all types have been bombarded with an ever-increasing tide of such models (and I have sampled only a small portion in the previous section!). It is time to stand back and assess the relevance and success of this build-a-model marathon. First, let us look at

what some evaluators have said about their colleagues' efforts.

Many professional evaluators are skeptical of this model-building trend. Early in the game, Cronbach (1963, p.672) stated: " ... I am becoming convinced that some techniques and habits of thought of the evaluation specialists are ill suited to current curriculum studies." Cronbach hit the evaluator himself, while Stake (1967, p. 524) aimed his pen at the people who should be using evaluation specialists: "The issue here is the potential contribution to education of formal evaluation. Today, educators fail to perceive what formal evaluation could do for them. They should be imploring measurement specialists to develop a methodology that reflects the fullness, the complexity, and the importance of their programs. They are not." Other indications of the failure of program evaluation are given by Guba (1969), Sorenson (1968, p.4), and Scriven (1967, p.53).

What are the symptoms of this failure of program evaluation in special education? The answer is simple enough: virtually no implementation in any area of exceptionality. Granted, the quality level of program evaluation might have risen slightly in federally funded programs because of the thrust for "increased accountability." But the real problem resides in the locally funded programs where such excellent opportunities for program evaluation exist. Here, there is almost no program evaluation at all. How many learning disabilities programs today are really being examined in a formal sense using systematically gathered data? Please note that I am not talking about program evaluation in terms of the usual indices of number of dollars spent, number of certified staff, number of children served, etc. Rather, I am talking about formal statistical evidence of any gains made in the program as a whole, derived of course from the gain data on individual pupils. I am talking also of comparative gain data of one learning disabilities program pitted against another program presumable aiming at the same goal but with

different means. This type of data just is not being provided on a routine basis for decision-making in locally funded, on-going programs. About the only evidence of formal program evaluation that is visible lies in isolated occurences of university research project evaluations in local schools (just examine any professional journal!).

So much for the obvious symptoms of the failure of program evaluation. What are the underlying causes? I think one primary factor has been the emphasis on mass dissemination -- in professional journals, conventions, etc.-- of the general program evaluation models. A few years ago, during the truly primeval stage of development in program evaluation theory, there were virtually no generalized guidelines to follow. Thus, initially, models such as CIPP and EPIC performed an admirable service in causing awareness, to some degree, in program administrators (but not in teachers!) of the need for program evaluation and of what its basic features are. However, with the flood of literature in this field (books, monographs, articles, speeches), I think the administrators and teachers in the field became disillusioned. After the initial dissemination of the models, nothing new was being said. There is an even more basic flow in the massive, never-ending, model-building epidemic: lack of specific advice within the models for actual implementation of the evaluation. If any learning disabilities educator examines the several program evaluation models presented earlier, he will probably remark: "I understand what you are saying, and it all seems very logical. <u>But</u>, I know I myself will never be able to use the model in my particular situation because no really specific guidelines are given. I would need expert evaluation help to use the model but do not know where to get such help. So I will forget the whole thing!" And there is the crux of the matter as I see it. The models have reached their level of functional incompetence with respect to the real world.

This problem of functional incompetence of models is basically one of analytical overkill. In the past few years, the program evaluation models have been refined, re-refined, ad nauseam. Indeed, some evaluators have even turned t eir by now finely honed analytical skills to a higher level of synthesizing: "meta-evaluation" and "taxonomies" of evaluation designs (cf. Scriven, 1969, and Worthen, 1968). Whether these super-analytical efforts be worthwhile or not, we had better slow down and re-examine our position, evaluators! The educators in the field have been left behind! Clearly, model building is not having a very salutary effect on education. In regard to the model-builders, Finn (1969, p.18) asked: " ... is it possible that they have, in fact, over-analyzed the process ... . Have, in fact, these analyses departed from operational reality, at least in the sense that the practitioner would not know what to do with them?" Thus, Finn suggests (p.19) that perhaps program evaluation has acquired that dreaded affliction known as "hardening of the categories!"

Let us also ask at this point in the model-building game just for whom the recently developed models are intended. We know they are not meant for the program administrator or teacher; they cannot handle the model on their own. What about professional evaluators? Could the models be aimed at increasing their competence? I think not. After reading the initial CIPP and EPIC models years ago, I as an evaluator have not received any new insights from the spate of publications issued since that time. The models appear to be stimulating thought in no one. They are highly repetitious and are probably doing more harm than good at this point.

In other words, the dissemination function of these models has outlived its usefulness. It is time to put the models on the historical section of the educational bookshelf. In my opinion, a model is supposed to lend a unique perspective not ordinarily realized by the majority of practitioners

in the field for which the model is meant. The models did this years ago, but no longer do.

A second major cause for lack of widespread implementation of program evaluation has been the absence of an aggressive "sales campaign" by existing agencies: county offices or intermediate units, regional materials and/or resource centers, universities, and state departments. Educators are traditionally slow to adopt innovations. Thus, some vigorous prodding is needed. Existing agencies that have program evaluation consultation capabilities must assume the responsibility to keep knocking on the doors of potential clients. Simple advertising of the availability of such services is not enough.

As a solution for those in need of professional evaluation assistance, one might as this point, suggest that the answer is simple: go to the local university evaluation service bureau. However, how many educators in the field <u>really</u> would feel free to call on consultants at universities? Not very many, I am afraid. There is an inherent distrust of universities in many educators. Some might even say: "The only consultation we ever received was a request to do research in our school; we never got any practical benefits from it. Any program evaluation consultation we get will probably be equally impractical! So why bother?" Perhaps this attitude is unjustified on the part of educators with respect to some of the more service-oriented universities. However, the attitude does exist, and it must be coped with.

One might also suggest that the program administrator obtain consultation from an agency like EPIC. This is fine for those who live in the vicinity of Tucson, Arizona, or near a handful of similar agencies. However, the vast majority of learning disabilities educators must do without such consultation, and thus without program evaluation itself.

A much more powerful solution is needed. Before suggesting a possible resolution of this sorry state of the art, let us examine briefly a typical

example of learning disabilities program evaluation. Perhaps too much has been expected of formal program evaluation. Let us see just what an evaluator might be able to deliver.

CHAPTER II

ILLUSTRATION OF FORMAL AND INFORMAL PROGRAM EVALUATION IN A LEARNING DISABILITIES CONTEXT

A brief example of some general aspects of program evaluation in a learning disabilities program has been given earlier in connection with the CDPP-CIPP model. To make that example more specific, let us assume that a group of thirty dyslexic children have been diagnosed as having comparable etiologies, that they lie within a relatively narrow age span, and that other pertinent factors are comparable among the children. In other words, meaningful comparisons can be made among various subgroups of the children. We will also assume that concrete action has been taken to carry out the preliminary phases of the CDPP-CIPP model. For example, during the context evaluation phase, a diagnostic pretest of reading deficit has been given to all children. Using this information and other data from each child's records, needs have been determined. Since some kind of perceptual motor training program was considered appropriate, specific measurable objectives were specified in both the perceptual motor and reading achievement domains of behavior; each objective was to be measured by a corresponding standardized (or, if more appropriate, locally devised) test. Before we enter the scene, let us also assume that the design or input evaluation phases have been partially accomplished in that alternative training programs have been examined, all within the light of the constraints of the school. This sets the scene for the example to be discussed below. All of the above steps have been accomplished by an evaluation consultant working with the program administrator and staff.

As we enter the scene, the program evaluation is ready to conclude its design or input evaluation phase. The evaluator must now decide what type of data gathering scheme would be appropriate. It has been agreed among all involved in this planning that the pretest of reading deficit can be used to divide the pupils into three groups of ten each: minimal deficit, moderate deficit, and severe deficit. All agree that specific information on the ways in which these three broad classifications of dyslexic children progress throughout the perceptual-motor training program would be valuable for decision-making on a short or long-term basis. Since the program will be run during the full academic year, it must be decided how many tests to give during the year. For purposes of in-process quality control, it was decided to give three middle-of-the-year tests as well as pre- and post-tests (all testing occasions use the same tests, or better yet, parallel forms of the same test). The resultant data collection scheme is given in Figure 2. The evaluation consultant concludes the design or input phase by specifying the type of statistical analysis to be used on the data: in this case, perhaps a "repeated-measures analysis of variance." It should be noted here that complicated statistical methods should never frighten program administrators or teachers away; the evaluation consultant has primary responsibility for selecting, performing, and interpreting the analysis.

At this point, one might ask how such a formal evaluation can aid both the program administrator and teacher in reaching rational decisions. In Figure 2, I have sketched in average learning curve profiles for each of the three diagnostic categories: minimal, moderate, and severe. Before entering into any detailed discussion, the reader should note that the collection of test data during the three in-process testing occasions (1/4, 1/2, 3/4) constitute the process evaluation phase of the CDPP-CIPP model, while the post-test comprises the product evaluation phase.

FIGURE 2.

SIMPLIFIED EXAMPLE OF PROGRAM EVALUATION DESIGN

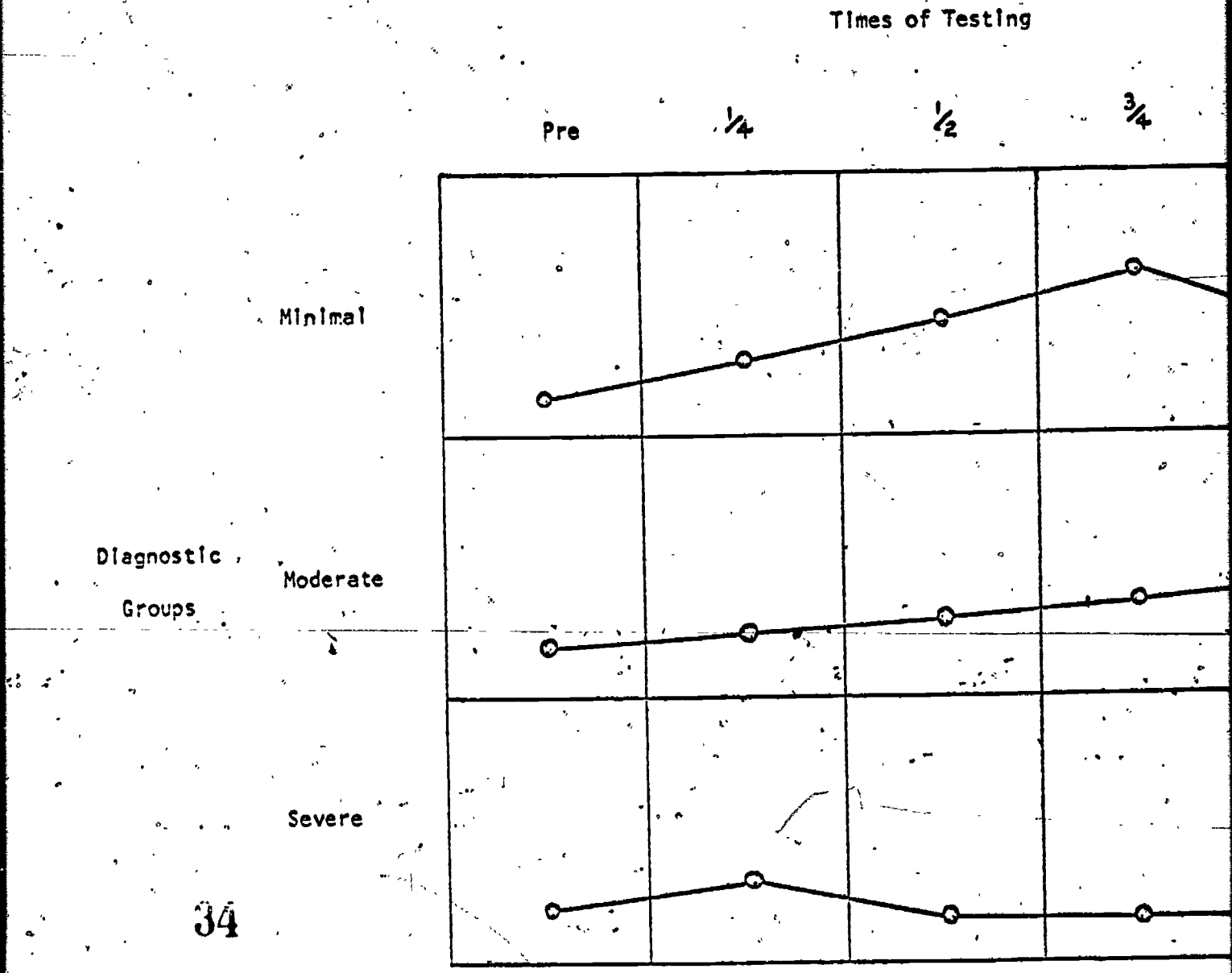
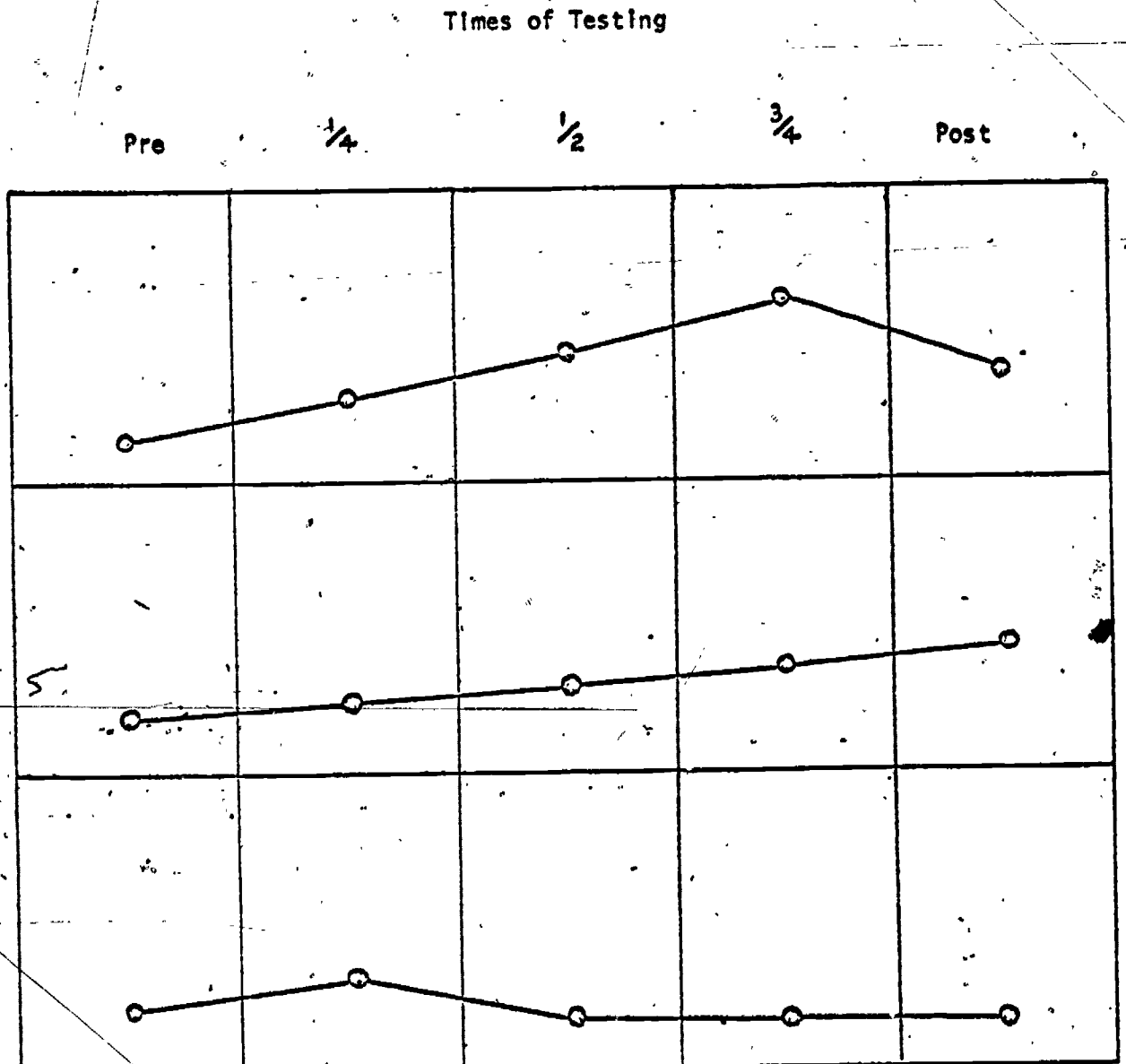


FIGURE 2.

SIMPLIFIED EXAMPLE OF PROGRAM EVALUATION DESIGN



Let us first consider the benefits to be gleaned for the program administrator from this formal program evaluation. He may note that the minimal reading deficit group gain nicely throughout the four quarters of the year until the final testing, then drops off. If the administrator examines the programming approach used with the "minimal" group, he might be able to isolate some of the probable causes of this change for the worst. Similarly, if the administrator considers the profile gain curve shown in the bottom row for severe-deficit children, a large drop-off in remediation occurs after the second quarter of the year; this in-process measure would tell the administrator to make some on-going changes before the end of the year. Granted, group profiles, averages, and so on, which are the working tools of formal program evaluation, have deficiencies (e.g., covering up finer differences among pupils). However, for situations in which more or less common features of remediation (even if individually administered) are applied to certain types of children under the broad heading of dyslexic, formal program evaluation can yield valuable benefits for the administrator.

What about the teacher? Even in such a necessarily oversimplified example, informational benefits aimed toward remediation should be evident. The teacher will play the major role in the data gathering process and will be making immediate, day-to-day programing changes (process evaluation). He will maintain a score sheet like the one in Figure 2 but subdivided into additional horizontal rows within each of the three diagnostic categories already shown. Each child's name will be appropriately placed along the left of the data maxtrix in the correct diagnostic category. The scores of each child would be placed on his small, horizontal slice of the matrix. The teacher might want to keep individual gain profile curves on each child while gathering the data for the program administrator and evaluator. In this way, the teacher would be able to see at a glance whether or not an individual child's

remediation was having a beneficial effect, and to act accordingly while there is still time to make a meaningful change in the programming for the child.

At each major point in the data collection process, the data sheet is copied from the teacher and fed to the program evaluator to suggest and arrange for appropriate analyses. These global program evaluation findings (that is, averaging and pooling to gauge the progress of the general types of children as a whole) are given to the program administrator. Thus, ideally, both the teacher and administrator would obtain appropriate feedback for decision making immediately. Decisions can be made during the program's operation and at its end (product evaluation).

It must be remembered that this is just one simple example of a program evaluation design. The sophistication of the formal analysis and feedback increase according to the desires of the administrator and the flexibility of the program itself. Of course, the ultimate success in terms of utility of any formal program evaluation depend on the willingness of the administrators and staff to use the findings in an intelligent way. Formal program evaluation does have limits (Stake, 1969; Wardrop, 1969). A great deal of debate has also centered around the differences between program evaluation and tightly controlled research (Schalock, 1970). No one will deny that evaluation studies lack a great deal of experimental control in the purist's sense of the word. However, if formal program evaluation is coupled with informal program evaluation, an intelligent basis for making decisions arises.

How do informal program evaluation methods enter the picture? Teachers and other staff members are continuously making use of these techniques when they administer their "homemade" tests, construct anecdotal records on individual children, on-the-spot observations of emotional difficulties, estimates of ability to interact with classmates, and so on. Too often these

types of data are shunted aside as being "nonstandardized", "subjective", etc. Why is it that teachers -- using their informal, subjective assessment techniques -- often come up with a much more effective remedial prescription for children than do objective, outside "experts" with all their standardized testing instruments? Also, how many federally funded programs for thd disadvantaged or other exceptional populations have been judged dismal failures in terms of standardized test data alone? Formal testing evaluation is quite limited at times, and for this reason alone informal data gathering procedures should be used wherever appropriate to obtain a more complete picture of what is occurring in a program. Karl (1970); Reynolds (1967); and Kunzelman (1969) have stressed the great potential of informal teacher evaluation. Clearly, both formal and informal evaluation procedures are needed in any serious assessment effort.

CHAPTER III

GENERAL IMPLEMENTATION STRATEGIES OF FORMAL PROGRAM EVALUATION (NORM-REFERENCED MEASUREMENT)

Thus far, many grandiose schemes have been advanced for carrying out program evaluation concepts in a learning disabilities context. But who will be available to the program administrators and teachers for providing custom-made evaluation consultation? Most existing general service agencies at the district, county, regional, and state levels do not have full-time program evaluation consultants. And I hope I have demonstrated that the evaluation models are far too general to be of any real help for specific program evaluation problems. We have also discussed why universities probably will not be asked to provide consultation in this area, I would like to propose a new type of general service agency that might form part of the answer. It is time to stop building models and start building consultation agencies.

The major thrust in any attempted resolution of the poor quality of existing program evaluation in learning disabilities, in my opinion, must lie in providing custom-made evaluation consultation to any qualified professional in need of it. Ideally, I would suggest that program evaluation centers be set up in strategic locations across each state. However, I realize such schemes are not always practical, and some compromise must be found.

There are two main avenues that appear feasible. First, county offices could hire one or two program evaluation specialists. Such people would have training at least at the Master's degree level in educational research and measurement. Thus, the county office could provide custom-made consultation, not only to the various exceptionality programs run by the county but also to

those special education programs run on an individual school district basis. In fact, the evaluation specialists could probably also handle program evaluation consultation requests from regular educators in each of the individual school districts in the county. The entire educational community stands to benefit.

If for political reasons or otherwise, it does not seem likely that a county service unit will become program-evaluation-minded, then regional general service agencies would have to be staffed with program evaluation specialists. For example, a large number of federally-funded instructional materials/media/resource centers have sprung into operation during the last few years. These centers usually serve large but still realistically sized regions. The provision of individualized program evaluation consultation services would be easy to append to the existing operations. Hopefully, both the county evaluation units and the media/materials/resource center evaluation units would have state sanction, encouragement, and even funding. The expenditure for salaries and operating costs of the two or more evaluation specialists in either type of agency would be negligible compared to the benefits which could be reaped in program improvement.

One cautionary statement of policy must be advanced, however, from past experience in such ventures. It is quite clear that many ongoing programs -- perhaps even the majority -- will not make use of a service even though it is announced as being available, free, and sophisticated. Program administrators and teachers have seen too many gimmicks and "revolutionary ideas" come down the road in recent years. Thus, program evaluation consultation services must be sold. It is the responsibility of the evaluation specialists to undertake a vigorous advertising campaign (brochures, monographs, on-site visits, telephone calls, personal letters, etc.) to stir up individual consultation requests. It is understandable for a program not to want to involve itself

in more "paper-pushing" than at present if possible; to many educators, the regional or county evaluation unit appears to be just one more example of bureaucratic entanglements. Such negative images must be offset through proven performance.

Besides actively soliciting evaluation consultation business from the programs in its service region, the evaluation unit should also conduct program evaluation workshops that serve as a dissemination function of the agency. Here is one rare instance where the general program evaluation models can still be of some value to the uninitiated. A small number of clients would participate in the workshop. The subject matter might consist of simulated evaluation exercises in learning disabilities or of back-and-forth discussion of actual clients problems.

The main service of the evaluation units would be to offer individualized, custom-made program evaluation consultation on demand by any client. However, the agency would be remiss if it did not engage in information retrieval and dissemination in program evaluation. For example, in the "design" or "input" phase of the CDPP-CIPP model, the final program of remediation must be decided upon in the light of competing approaches. How does one obtain information on all these competing brands of treatment? The county or regional service unit could house an information collection of research journals, ERIC, government publications, professional books, curriculum guides, technical reports, etc. Any client in the service region would be allowed to phone or write in a request to the center for a sophisticated literature search of all relevant findings in the area of concern. Also, in the selection of appropriate testing instruments, comparative information on prices, technical qualities, etc., could be provided. The evaluation agency would also be responsible for disseminating information on existing guides to program evaluation (e.g., Annas and Bowd, 1966; Grobman, 1968; Center for Instructional Research and Curriculum

Evaluation, and Cooperative Educational Research Laboratory, Inc., 1969; Meierhenry, 1969; Ahr and Sims, 1970; Mosher, 1968).

Let me conclude my "grand scheme" by throwing out a few words to those who may not agree with these ideas. A lot of potentially valuable schemes in learning disabilities die shortly after birth because of too much talk and too little action (indeed, an analogy may be made with the case of program evaluation models in all areas of education). The above "solutions" to the program evaluation dilemma in learning disabilities are, to me, rather obvious. We do not need a lot of local and state committees to conduct "studies" of the problem. All one needs is a few key people who can get things moving and keep them moving. The above ideas -- all of them -- have already proved effective in realistic, ongoing practice. There is simply no longer any excuse for the sad state of program evaluation in the field of learning disabilities!

Before leaving the realm of personalized program evaluation consultation services, a few words about the role of the evaluation specialist would be appropriate. I feel that, with occasional exceptions, fairly sophisticated statistical-inferential evaluation schemes can be applied to most learning disabilities programs in operation. Each program evaluation scheme is highly unique and usually applicable only in a narrow range of situations, before the evaluator has to shift gears entirely and devise a different design. I also want to dispel the myth that the program evaluator is, or should be, a "man-of-all-seasons" with respect to the whole range of educational technology. Most of the recent breed of evaluation specialists are usually competent only in the fields of statistical analysis, design methodology, and test construction and use. These specialists are not experts in curricular philosophy and thus cannot and should not make value judgments about remediation planning. If program evaluators are being honest withthemselves, I seriously doubt whether they can pronounce judgment on a program, other than to yield some inferential

data on the quality of the intermediate and final products of the remediation and to suggest possible interpretations to the program personnel. Only if one can find an expert curriculum specialist "retreaded" into an evaluation expert (and I mean fully retreaded!) can the program personnel expect to have ultimate value judgments about their program made for them by the evaluator. Almost without exception, the program administrator and his staff members must make the final value judgments about the program. I also want to make clear that I am not asking for programs to be unrealistically twisted into highly sophisticated research projects. This would be the usual criticism against one who emphasizes as much formal design methodology as possible in a given situation. All that I am advocating is that the field practitioner and program evaluator join heads in coming up with the most sophisticated evaluation design possible for the particular project in question without project distortion.

CHAPTER IV

RELATIONSHIPS OF FORMAL PROGRAM EVALUATION TO EXISTING STATEWIDE AND NATIONAL ASSESSMENT SCHEMES

All of the discussion thus far has emphasized custom-tailored program evaluation schemes. In one case, for example, a perceptual-motor training program might employ the Southern California Perceptual Motor Tests, while a similar program in a different region might administer the Frostig Developmental Test of Visual Perception. It is difficult to compare the results of one program evaluation with those of another, if not impossible. It has been tacitly assumed that the results of any given program evaluation scheme are useful only to that specific program. Could a more generalizable program evaluation scheme be achieved for almost all programs in all areas of exceptionality? In other words, could comparable program evaluation schemes be devised? Current activities in regular education indicate the answer is "yes." There are two main facets to this issue: (a) statewide assessment, and (b) national assessment.

Several states have initiated statewide assessment or evaluation schemes. In general, a group of subject matter experts and others has agreed upon a series of measurable objectives in the various domains of student behavior that any regular educational program would hope to achieve. A series of tests is found or devised for each major objective. Schools of various types of specified characteristics (such as pupil population size, community size, geographic location, etc.) are sampled randomly. The same battery of tests is administered by local personnel in the selected schools. From such test data, score distributions and norms are derived. Finally, individual schools get feedback on how their students compared with similar (and dissimilar) students across

the state; manipulatable characteristics of the schools that appear to be highly related to ongoing pupil behavior are also identified (such as academic preparation of teachers, per capita expenditure, etc.). In most statewide assessment efforts, the battery of tests is administered only once a year in only a few grades; no gain data is gathered.

Dyer and Solomon (1970, p.4) have stated: "Ultimately, we need to be able to answer the question: What educational processes work in what kinds of schools for what kinds of kids?" One must remember, however, that these pilot efforts have been initiated only in the realm of regular education; special education, in most cases, has not even been touched. O'Reilly (1970, pp.3-4) describes New York's statewide assessment program: ,"Each fall, all public and nonpublic school pupils in grades 1,3,6, and 9 have received certain standardized tests: a readiness test for grade 1 and tests in reading and arithmetic for grades 3, 6 and 9 ..." However, O'Reilly does not feel a once-a-year data collection is adequate for program decision making at the state level; he suggests that more data collection points be inserted into the course of a year. Among other things, meaningful gain data can thus be generated. One can see the analogy with the custom-tailored program evaluation example mentioned earlier with respect to gain analyses.

Loadman and Major (1970) have described Michigan's statewide assessment efforts. Educational Testing Service (ETS) of Princeton, New Jersey, helped construct tests to measure program objectives considered suitable to the two grades selected for assessment: 4 and 7. Besides providing each school building within a single school district with results, more general results will be given by a two-way classification of community type (5 types) by region (4 regions). Other analyses will also be performed.

The Bureau of Educational Research of the University of Virginia has been working with the Virginia State Department of Education since August, 1969), in

one of the newer statewide assessment efforts. Woodbury et al. (1970, p.7) says: "Specific behavioral objectives ... include English (literature, language, composition), Mathematics, Reading, Science, Social Studies as well as personal and social categories of affective behavior. More general behavioral objectives were developed for Foreign Language, Health and Physical Education including psycho-motor skills, Vocational Education, Early Childhood Education, Work Study and Library Skills, Special Education, Art and Music."

Other aspects of statewide assessment have been described by Kearney (1970), Michigan Department of Education (1969), and the Pennsylvania Department of Education (1968).

I have mentioned such statewide program assessment or evaluation efforts in the hope of stimulating learning disabilities educators and other special educators into thought about devising a similar model in their respective domains. The possibilities are exciting or frustrating, depending upon one's view of statistics and testing. Will there be problems of major proportions in adapting such schemes to special education? Most certainly! For example, each area of exceptionality will probably have to be treated separately. The physically handicapped cannot be expected to take some physical performance tests, while the severely retarded will not be able to wade through all but the simplest conceptual achievement tests. It is my hope that learning disabilities educators will at least try to adapt some of the ideas of statewide program assessment for their own area.

However, one need not stop at the state level in the attempt to devise a "standardized" program evaluation system. The National Assessment of Educational Progress (NAEP) has been underway for a few years in regular education. This effort began in 1964. Since the ideas are basically the same as in some of the statewide assessment schemes, the reader can refer to the large body of

literature on the subject (Saylor, 1970; Katzman and Rosen, 1970; Groff, 1970; Womer, 1970; Findley, 1970; Katzman, 1970; Caps, 1970; Ebel, 1970).

CHAPTER V

USE OF CRITERION-REFERENCED MEASUREMENT IN FORMAL PROGRAM EVALUATION, IN DISTINCTION TO NORM-REFERENCED MEASUREMENT: REVIEW OF LITERATURE

Introduction

The first four chapters of this monograph have considered the use of norm-referenced measurement in conducting formal program evaluation. In other words, standardized tests are used in accord with accepted research theory. Such an evaluation strategy is quite appropriate when only a global overview of an on-going program is desired. A classical research strategy is used which would have at least a pretest and a posttest, and preferably one or more equispaced measures during the in-process part of the program. However, there will no doubt be special projects with which the Commonwealth's Bureau of Special Education will be connected and for which the usual classical research evaluation design will not be adequate. Such situations lead one to a much more intensive type of formal program evaluation known as criterion-referenced measurement. A case in point with which most members of the special education staff throughout the Commonwealth will be able to identify is the National Regional Resources Center of Pennsylvania (NRRC/P). Here is a major project that is linked directly to the state Bureau of Special Education, as well as to regional special education agencies in the central and eastern parts of the state. Another aspect to criterion-referenced measurement that will be discussed in another chapter is data-banking activities. For ease in discussion, the following abbreviations will be used:

norm-referenced measurement (NORM), criterion-referenced measurement (CRM), and data banking (DABA).

All material in these next few chapters that pertain to NRRC/P, CRM, and DABA was produced in connection with articles Lester Mann and Bart Proger are writing for project dissemination purposes with NRRC/P. The materials contained herein have been modified so as to tie in directly with the many aspects of the total formal program evaluation model presented in this monograph.

Consider a child who has been referred to NRRC/P as having specific reading disability to the extent that he cannot function at even a first grade independent reading level. Suppose further that as part of the psychoeducational programming for this child that one specific objective in picking up the child at his current level of functioning and carrying him forward, is to have him recognize letter differences among vowels embedded C-V-C trigrams. Presumably, during the initial referral process, this child has already been diagnosed as having a deficiency in this particular reading skill area. Further, other components of the reading process will have been similarly diagnosed to provide some rough basal guidelines of where the child presently stands. However, it must be emphasized that no undue weight will be given to basal functioning levels. Rather, the emphasis will be on what final levels of functioning the child achieves. This measure is what really constitutes the pay-off evaluation of success.

True, in a tightly controlled experiment one is interested in pre-post differences within and among treatments -- the statistical significance phenomenon. With respect to the real world, however, many researchers have been questioning the legendary thrust toward significance. We need a mode; different from the usual experimental one to answer the types of practical questions that NRRC/P is asking. As mentioned previously, the project wants to answer the frequently asked but as yet unresolved questions of: (a) how much

success can be expected in certain specific skills associated with selected subject content areas as taught by a specific approach "A"; (b) how long it took a certain approach "A" for teaching that skill to reach the observed level of success in (a); and (c) how the answers for questions (a) and (b) for approach "A" compare to competing approaches "B", "C", etc. Because the psychoeducational programming thrust of some components of NRRC/P demand that programming recommendations be made in terms of a highly specific subject content analytical breakdown of the total task into its subskills, the usual standardized test, classical evaluation design is not appropriate.

Thus, NRRC/P has decided upon the use of criterion-referenced measurement, with overtones of achievement monitoring and data bank activities. Popham and Husk (1969, p. 2) have given one interpretation of criterion-referenced measurement (CRM): "It is not possible to tell a norm-referenced test from a criterion-referenced test by looking at it. In fact a criterion-referenced test could also be used as a norm-referenced test -- although the reverse is not so easy to imagine. ... At the most elementary level, norm-referenced measures are those which are used to ascertain an individual's performance in relationship to the performance of other individuals on the same measuring device. ... Criterion-referenced measures are those which are used to ascertain an individual's status with respect to some criterion, i.e., performance standard. It is because the individual is compared with some established criterion, rather than other-individuals, that these measures are described as criterion-referenced. ... We want to know what the individual can do, not how he stands in comparison to others."

Nonetheless, Simon (1969, p. 259) cautions CRM advocates not to get carried away in the wash of jargonese: " ... strictly speaking the distinction between criterion-reference and norm-reference applies not to the test but to the test scores. In other words, the distinction does not relate to the nature

of the test or to the content or form of the items, but concerns primarily the interpretation and use of the scores from the test. It is perfectly appropriate for a single test to report both absolute performance (criterion-referenced) scores and relative-performance (norm-referenced) scores."

While Simon is technically correct, nonetheless, NRRC/P will be forced by the very nature of its objectives to make a working distinction between NRM tests (standardized ones, i.e., those with norms) and CRM tests (custom, project constructed tests). Getting back to the example at hand of the child getting training in recognizing vowel differences embedded in C-V-C trigrams, a CRM test would be constructed for the measurement of degree of success at the end of the week-and-a-half (or whatever) unit of instruction. The measurement experts on the NRRC/P staff would construct the CRM instrument to become a part of the achievement monitoring system (AMS) for this child.

The advantage of CRM testing is that the project personnel decide what the criterion of degree of success should be for a child with disabilities such as the present subject exhibits. Perhaps for this particular CRM test of various types of C-V-C trigrams, the NRRC/P staff will decide that 65% competency is needed before the child is allowed to move on to the next sequential area of subject matter. For a more crucial subskill area, perhaps 85% competency pm the CRM test will be demanded. Flexibility, realism, and practicality are primary attributes of the CRM system. For the better part of this century, special educators have been guessing at the answers to questions such as (a), (b), or (c). Other than a few isolated experiments in often contrived environments or rather loosely conducted demonstration projects, the answers to such questions have gone begging. Hopefully, the NRRC/P, through its CRM-AMS, will begin to build a data bank (DABA) from which future educational researchers and practitioners can draw.

It should be noted also that the results of the last few years of federally

funded projects will be utilized to their maximum potential in establishing and operating the CRM-AMS system. For example, for the purposes of breaking the sequential arithmetic curriculum into its components and for gaining programming ideas, project PRIMES will be utilized. Further, projects that are generating program materials along lines of a sequential task analysis/behavioral objectives basis will be contacted as sources of materials. In terms of specifying behavioral objectives and developing CRM test instruments, the Instructional Objectives Exchange (IOX) housed with the Center for the Study of Evaluation at UCLA will be tapped wherever appropriate. (see Skager, 1970).

For years the main thrust in educational measurement was away from teacher-made tests and towards standardized instrumentation. No doubt a large causative agent in this trend was the great volume of ever-increasingly sophisticated educational research studies, which usually emphasized standardized tests and rating scales. The "home-made" or locally produced variety of test was somehow frowned upon and judged useful only in granting report card grades but never for whole-year or global-program evaluations. Further, if CRM tests are to be used quite frequently as an in-process type of quality control at the ends of major units or blocks of instruction in the subject matter sequence, then by the very nature of this frequently occurring measurement task, CRM tests must be custom-built to the users requirements as the measurement needs arise. In other words, what we are saying, curiously enough, is that "home-made" testing instruments are back in vogue but -- more importantly -- are also back in respect, when used intelligently and legitimately. This almost circular historical trend in measurement methodology is indeed strange but -- in education -- not surprising. Because CRM is rather new in the field of special education, a review of the literature in this field will be helpful in understanding part of the function of NRRC/P. Also AMS and DABA literature will be covered for the same reasons.

One must realize the jist of what is being proposed here. A national project is considering the use of home-made tests (Albeit in the refined vein of CRM) to answer some research questions of high priority in the LD and EMR fields. Cannot this enterprise be questioned on the grounds that home-made tests -- even of the "higher" CRM variety -- still have the often-cited flaws of "looseness" in measurement methodology? Are not CRM tests still plagued by subjectivity and possibly by lack of adequate reliability and validity? Klein (1970, p.3) has raised some of these questions, and his arguments merit serious consideration: "The ... use of criterion-referenced measurement would be a laudable practice if one knew how to determine what criterion objectives to specify, or what level of performance constitutes their attainment, or how to interpret the results if the objectives are or are not achieved. To illustrate this point, let us suppose that a new course unit in 10th grade biology let to 30% of the students attaining all of the unit's 20 objectives, 50% of the students attaining 15 objectives, and only 20% of the students achieving less than 10 objectives. These results look very impressive and a school official might be very pleased with the effectiveness of the program. But would he still be happy if he discovered that most students could achieve 10 of these objectives before taking the unit, or that the criterion of attainment was 1 out of 5 items correct per objective, or that the items used to measure an objective were not truly representative of the range of items that might have been employed, or that 80% of the students at other schools (having students of comparable ability) attained all 20 objectives using a criterion of 4 out of 5 items correct per objective?" (p.3)

Klein (1970) goes on to propose an eclectic test construction model based upon both CRM and NRM procedures. In effect, he is aiming his comments at standardized test producers and hopes that they will begin to issue instruments that embody the best features of both CRM and NRM. The first step is to specify

objectives in operational terms. Klein recommends that each objective embodied in the test should have at least three items to measure it. This guideline can be used in a forward sense to determine how long the test will be, or in a backwards sense to determine how specific the objectives should be. The second step is to find test items for each objective. Not only should the items be representative, but they should also represent different difficulty levels within the objective. The third step is to find test items that tap related objectives. "The reasons for measuring these kinds of related objectives are that they (a) provide information about the unanticipated outcomes of educational programs, (b) indicate how close a program (or student) came to meeting or surpassing the objectives (a), and (c) show the level at which subsequent educational treatments should be pitched. (p. 4)." The fourth step is to give the test user for each objective measured by the test a score and its interpretation. "Donald Jones (or Program #3) got four of the six items correct on objective number 7 (addition of whole numbers less than 100). Approximately 80% of the other students in Donald's class did this well. Students of equal ability in other classes (or programs) only got one-third of the items correct which is typical of the second graders in this state (i.e., the median score statewide on this objective is 33% correct). (p. 4)." With respect to writing objectives in terms of difficulty levels and levels of intellectual functioning, Klein recommends such "atlases" as Bloom (1956) and Guilford (1967).

It should be noted that the IOX is now an independent, non-profit corporation apart from the UCLA Center for the Study of Evaluation which is directed by Dr. Marvin C. Alkin. The IOX is directed by Drs. W. James Popham, Eva Baker, and John McNeil. This is effective May 31, 1970.

Mayo (1970) has argued elegantly for the individualization of instruction by means of appropriate CRM measurement. He calls the practices "mastery learning" and "mastery testing." Mayo suggests that a new conceptualization of

mental ability is necessary if a true matching of instruction to a child's specific needs is to occur: "Rather than thinking of aptitude as a kind of ceiling, Carroll (1963) suggested that aptitude may be related to the amount of time necessary to achieve mastery. (p. 2)"

The "mastery model" described by Mayo (1970, p. 2) has five features: "(a) Inform students about course expectations, even lesson expectations or unit expectations, so that they view learning as a cooperative rather than as a competitive [illegible]-ise. (b) Set standards of mastery in advance; use prevailing standards or set new ones and assign grades in terms of performance rather than relative ranking. (c) Use short diagnostic progress tests for each unit of instruction. (d) Prescribe additional learning for those who do not demonstrate initial mastery. (e) Attempt to provide additional time for learning for th[illegible]se persons who seem to need it."

In developing CRM (or "mastery") tests, Mayo points out that the usual requirements for maintaining an average item difficulty level of about 50% no longer hold; instead, the scores of pupils will tend to cluster in a skewed distribution around perhaps an 85% difficulty level. Educators molded more or less along traditional test construction lines will be somewhat disturbed in that "mastery tests" will seem to be almost too easy for a large portion of the pupils. However, this is in line with the different conception of learning that CRM is based upon. Given enough time and individualization of instruction, pupils should be able to achieve the majority of objectives in basic skill areas. This is the premise NRRC/P is working under. "The few who fail the item show a clear deficit, and this feedback indicates need for additional remedial learning sessions and repeated testing until items are passed (p. 3)."

Cox and Sterrett (1970, p. 227) have proposed a model that combines the best features of NRM and CRM: "(a) a precise description of curriculum objectives and a specification of pupil achievement in reference to these objectives;

(b) the coding of each item on a standardized test with reference to the curriculum, and (c) the assignment of two scores to each pupil, one reflecting his achievement on items that test content to which he has been exposed; the other his achievement on items that test content beyond his present status in the curriculum or not represented in the curriculum at all."

One extensive application of CRM that has many implications for NRRC/P is the Comprehensive Achievement Monitoring Project (CAM) of Dwight W. Allen and William P. Gorth of the University of Massachusetts. O'Reilly, Schriber, Gorth, and Wightman (1969) have prepared a lengthy manual that documents the implementation of a complete CAM system. Gorth has had primary responsibility for developing this CRM-CAM design. In the introductory part of their manual, the authors state: "The CAM procedure focuses upon the evaluation of achievement by more or less continuous monitoring of student performance relative to specific course objectives. Unlike traditional evaluation procedures which generally involve testing of students on discreto units of material, CAM generates performance data on all course objectives ... (at several points in time throughout the instructional sequence). The procedure consists of a battery of parallel (or equivalent) test forms which contain items representative of the span of the entire course and which are administered to all students at frequent pre-set, equal intervals. Each form contains an equal number of items (related to the specific objectives of a course) for each instructional interval and each item is used on only one form. Items are assigned to test forms by random sampling techniques and each form is from 10 to 40 items in length. Through the use of the random sampling of test items, it is literally possible to test hundreds of specific objectives over a group of students. Each test form is similar to a final test for a course. As the course progresses, the student should be able to answer an increasing number of items on the test forms corresponding to an increasing number of objectives mastered. Each student

receives a particular test form only once. Over the duration of these tests makes it possible to sample performance on every course objective over a given group of students at every testing." One can think of each test form given throughout the course of instruction as representing a barometer, with increasingly more difficult objectives corresponding to gradations of degrees. The more success the student achieves, the higher the barometer registers, and these readings can be put into a trend analysis over the passage of time for each student. Thus, individualized instruction can be monitored quite intensively. It should be noted that while the CAM system was originally operated on the basis of group profiles as contrasted to individual profiles, NRRC/P will concentrate on the latter.

Mathematical models are usually rare in the field of education. It is one thing to develop a psychological theory for a phenonemon to a rigorous mathematical modeling process. Pinsky (1970) has done just that with CAM. Further, the CAM originators have gone so far as to provide canned computer analyses for processing all of the monitoring data (Gorth, Grayson, and Lindeman, 1969; Gorth, Grayson, and Stroud, 1969).

CAM has been tried out successfully and realistically in a number of different situations. While O'Reilly et al. (1969) have summarized the technical details of how to implement every phase of a CAM system in future instances, Pinsky (1970, pp. 45-68) has given judgmental evaluations of systems already in operation. Several pilot locations have been selected for CAM projects; Duluth, Minnesota (for two consecutive years in a high school); Kailua, Hawaii (for three consecutive years with 11th and 12th grade trigonometry and algebra, and for two consecutive years with 11th grade American history); Hopkins, Minnesota (for two consecutive years with 11th grade algebra); Portland, Oregon (for three consecutive years with 9th grade algebra). Thus, one can see that CAM has been tried with what are perhaps some of the most complicated

subjects.

Each test form is called a "monitor." While each overall operation of CAM (with the exception of Duluth) can be termed a success, Pinsky (1970) nonetheless points out some operational difficulties that one is likely to encounter. A parallel-form monitor is usually given to each child once every two weeks throughout a course. Often a child will not take a monitor at the time it should be taken. Sometimes teachers do not have enough time to make use of the feedback data, or, if they do have time, will not put such data to full use. Turnaround time for processing the monitor data either by computer or by hand may discourage some. However, the benefits seem to outweigh by far the disadvantages of CAM. If used cautiously, monitor feedback data can be used to program for the deficiencies of a given child, or, at a different level, to change the general programming for an entire group of children. Successful performance on monitors by certain students can allow them to go into independent study or to advance more quickly, rather than be branched back over poorly learned material. Further, the program gets out of the old rut of pitting student against student and makes an individual compete only with himself on whatever time schedule he feels he can handle.

The CAMP Project is one of the few intensive ongoing CRM systems that is operating presently. As such, CAM deserves a long hard look at just what the operational and organizational requirements are. The "Guide ..." of O'Reilly et al. (1969) gives such details.

Cox (1970)has examined some conceptual difficulties of CRM with regard to technical issues of test construction (reliability, validity, and item analysis). He notes a trend in CRM in that most applications have been in the domain of individualized instruction. Cox begins his discussion of the technical issues by distinguishing between CRM and NRM: "When an achievement test is constructed as a norm-referenced measure the test items are written or selected

to maximize differences between individuals. Maximum discrimination is desirable to obtain the variability necessary for ranking individuals." However, Cox (1970) goes on to describe item analysis techniques from an earlier study (Cox and Vargas, 1966) that might be more appropriate for CRM." Two discrimination indices were computed for items on tests which had been administered both as pre and post-tests. The question of interest was the extent to which the two methods of item analysis yield the same relative evaluation of items. One index was computed using the common upper minus lower groups technique, thus providing information on how well each item discriminated between their groups. The second index involved both the pre and post-test and was computed by subtracting the percentage of pupils who passed the item on the pre-test, from the percentage who passed the item on the post-test. This index provided discrimination information between pre and post-test groups, indicating items useful for pre-test diagnosis. Results of the comparison between the two indices indicated that some items which are highly desirable for the pre-post test discrimination would be discarded by the typical item selection techniques, because they fail to discriminate among individuals taking the test. It was concluded that the pre and post-test method of the item analysis produced results sufficiently different from traditional methods to warrant its consideration in those cases where score variability is not the concern, such as in criterion-referenced measures." In terms of diagnostic procedures in special education, the pre-post CRM item analysis techniques seem to hold a great deal of usefulness. Cox (1970) concludes his examination of CRM technical issues by suggesting that perhaps the usual coefficients used to measure reliability and validity might not be appropriate because of the lack of enough variability.

The idea of using an achievement monitoring system in special education is not entirely new. Kunzelman (1968) described what he termed "data decisions."

(However, to our knowledge, the use of a monitoring system in special education that makes rigorous use of CRM in a legitimate way is new.) Kunzelman wants educators to engage in "self-help teaching," such as has been developed by the Experimental Education Unit of the Mental Retardation and Child Development Center at the University of Washington. Basically, Kunzelman's system consists of recording both correct and wrong rates of response for children within a teacher's class for a given content subject area. For example, if a certain teacher has been having particular trouble in getting one of her student to master a certain concept in arithmetic, she might decide to use a somewhat different tactic of individualized instruction than she had been using. To determine the relative effectiveness of the old and new approaches with that child, the teacher would have to maintain both correct and incorrect rates of response for a few days in arithmetic both before and after the point at which remediation tactics were changed. However, the hairy problems of just how much behavior to sample, when to sample, how to sample, etc., are not discussed by Kunzelman. These are precisely the issues met head-on by CRM such as Project CAM of the University of Massachusetts.

Emrick and Adams (1970) have provided what appear to be sounder cut-off points for making a "success-failure" determination on CRM tests. They use as their examples situations from the Individually Prescribed Instruction Project (IPI) of the Learning Research and Development Center at the University of Pittsburgh. Because IPI makes heavy use of CRM, such new mathematical models as Emrick and Adams propose are highly relevant to NRRC/P. The authors state: "IPI currently maintains an 85% correct minimum as a mastery criteria for any skill test (of which there are over 400). Although this criteria does have intuitive appeal, there is no convenient analytical or empirical justification for it. In particular, just as various skills may differ in level of difficulty in terms of mastery, so also might the optimal performance criteria in the test

situation vary. It may easily be that for some skills, a test score of 60% is indicative of mastery, whereas for others a score of 90% or higher would be required. In short, the issue is not whether a criterion referenced testing procedure is or is not appropriate to IPI, but rather how and at what level each criterion should be set." Emrick and Adams go on to propose a Bayesian algorithm for determining "success-failure" cutoff points.

Lundin (1970) has considered the role of CRM as a means to process evaluation of curricular materials still in developmental stages. He describes the experiences of the Minnesota Mathematics and Science Center Staff (MINNEMAST) in this regard. The research and evaluation team of MINNEMAST called their CRM system DRATS ("Domain Referenced Achievement Test Systems"). DRATS uses item sampling to extract the maximum amount of in-process information. While Lundin is in back of DRATS in particular and CRM in general, he warns: "if decisions based on sophisticated data do not result in improved student learning, then one can do without the luxury of sophisticated data until one develops sophisticated decision makers."

Several detailed descriptions of the CAM Project of the University of Massachusetts have been given recently (cf. Allen, 1970; O'Reilly, 1970, and Gorth, 1970). In particular, some key features of Project CAM in terms of CRM have been brought up by Allen, Gorth, and Wightman (1970). The authers state: "CAM measures achievement in a systematic way throughtout a course in the secondary or elementary school. It is comprehensive in two dimensions: 1. Time because achievement is measured throughout a course and 2. Course content because achievement is measured on all of the behavioral objectives specified for a course at each time. CAM uses several of the most modern techniques in educational measurement to obtain the goals it sets for reliability and validity. The techniques include item sampling which has recently been developed by Frederick Lord and longitudinal testing which has often been recommended to

measure change or growth. Both of these ideas have been tied to computer programs for rapid analysis and reporting of the results to students, teachers, and administrators."

Another major set of benefits derived by CAM, claim Allen, O'Reilly, and Gorth (1970), is that: "At each test administration, performance or objectives not yet taught is pretested, performance on objectives just taught is immediately post-tested, and performance on objectives taught earlier in the course is measured for retention."

Flexibility is another major virtue of the CAM system: "Monitors are indented to be short tests, perhaps ten to thirty items. Whether or not a single form covers all objectives for a course is a function of the proportion of objectives to items-perform. It may be necessary to randomly sample (without replacement) the objectives, before doing the same on the test items for each selected objective."

Allen, O'Reilly, and Gorth (1970) describe several different types of feedback, at either the individual or group level: "For individual students: After each administration: 1) total score on that and all previous administrations, 2) a graphic presentation of the above, 3) a right-wrong indication for each item on the monitor, coded by the objective represented. At the end of the course: 4) average scores, across all monitors taken, on items categorized by use into three groups -- pretest, immediate post-instruction and retention of varying lengths of time. For whole group or subgroups (e.g., one classroom, highest and lowest quartiles): After each administration: 1) percent answered correctly out of all items across all monitors, for each objective. Periodically, as desired (e.g., every 3-5 administration): 2) trend data, or achievement profiles, for total score and for each objective. At the end of the course" 3) same as number 4 under individual students, 4) item analysis (using whole group only), treating each item in three separate ways, by its three functions -- pretest, immediate post-instruction, and retention measure."

CHAPTER VI

A DETAILED DESCRIPTION OF A CRITERION-REFERENCED MEASUREMENT SYSTEM THAT WOULD BE SUITABLE FOR SPECIAL STATE-CONNECTED PROJECTS, SUCH AS THE NATIONAL REGIONAL RESOURCES CENTER OF PENNSYLVANIA

Introduction

From the previous chapter, the reader should now have a command over what the concept of criterion-referenced measurement (CRM) means and what some of its advantages and disadvantages are. While formal program evaluation in the norm-referenced measurement (NRM) sense is the most feasible route to follow in implementing a large-scale accountability system, for any intensive examination of exactly what is happening in special education classes, the author feels CRM is the only real answer at present. For these reasons, whenever special projects are run that are connected with the state Bureau of Special Education or, for that matter, are run strictly on the local level, a detailed description of how the projected CRM system will operate in the Eastern Suburban Division of the National Regional Resources Center of Pennsylvania (NRRC/P) is given here. It should be noted at the outset that the NRRC/P CRM system can be modified to accomodate the specific needs of any program.

BACKGROUND ON NRRC/P

The National Regional Resources Center of Pennsylvania began July 1, 1970, with one year of planning. NRRC/P is a cooperative effort that combines the

resources of several existing public special education agencies in the central and eastern third of the Commonwealth of Pennsylvania. The principal investigator of this federally funded, long-range projecr is Dr. William F. Ohrtman, Director of the state Bureau of Special Education in Harrisburg. NRRC/P is devoted to intensive study of the efficacy of various programming techniques used with learning disabled children of elementary school age (in the sense of the national definition of learning disabled). Several experimental classes are being established in urban, suburban, and rural areas in both the central and eastern portions of the Commonwealth. Rather than concentrate on global programming questions of a long-range nature (which program evaluation efforts tend to mask the more crucial features of why or why not a total programming technique was successful), NRRC/P will be looking intensely at how well small manageable units of instruction work with certain types of learning disabled children. For example, if one unit of instruction were to have as its primary objective the mastery of a certain family of words, then NRRC/P wants to know: (a) what level of criterion mastery can be expected over a specified period of time with programming approach A that is used to teach the family of words, (b) how the level of success achieved with approach A compares with approaches B, C, ..., (c) what different levels of mastery are possible with learning disabled children of different impairment levels under programming approaches A, B, C, ..., and (d) what cost-effectiveness factors enter into approaches A, B, C, ... At first glance, this list of questions would seem to be an over-ambitious project. However, the criterion-referenced measurement system of NRRC/P has allowed for the study of all these problems. This CRM system seems to hold several useful implications for <u>any</u> area of special education; let alone the learning disabled.

In dealing with learning disabled and minimally brain damaged pupils, individualization of instruction is of utmost importance. For this reason,

customized, psychoeducational programming will be used for each pupil. However, before devising an individualized prescription for each child, a modified form of small group instruction will be used at the start of the year to enable the teacher to spot those pupils who are in need of immediate, intensive, individualized help.

The basic program operation of the national project will be organized around two major components. First, the small group instruction will form the central component or track about which all individualized efforts will be oriented. The intent of the project is to keep the child in the main instructional track whenever possible. The small group instruction forms the regular education component of the project. Second, whenever a child begins to run into severe educational difficulties that cannot be handled within the regular small group mainstream, he or she will be sent to the resource teacher for one-to-one individualized help, or an itinerant resource teacher will try to deal with the child within the classroom.

The instructional sequence, whether in the regular, small-group mainstream or in the resource-teacher, individualized-prescription situation, will be divided into "instructional modules." Each module is used only over a relatively short period of time, perhaps two weeks. The instructional module is organized around a set of highly specific objectives stated in measurable behavioral terms. Each child's achievement both before and after going through the module is measured by "monitors," which are special types of tests (interpreted in the CRM sense). The pre-monitor is used before the module is entered upon, and the post-monitor is used after the child has completed the module. If the child demonstrates inadequate achievement on the post-monitor, he is then brought into contact with the resource teacher for one-to-one individualization of instruction. With the resource teacher's help, the child is again taken through the instructional module which he has not mastered in the regular in-

structional mainstream. After using similar forms of post-monitors to check again on the child's understanding of the module with which the resource teacher has given him help, a decision is ultimately made as to when to send the child back to the instructional mainstream.

THE CRM SYSTEM OF NRRC/P

To carry out such an instructional programming system, a corresponding measurement and evaluation system must be devised. While standardized tests will continue to be used as part of the usual individual psychological screening evaluation given to all children in the national project, the flexibility of such tests for measuring change within any given pupil is quite limited. First, any standardized test selected will usually embody only very global program objectives; the specific instructional objectives of a certain module will only be reflected occasionally within a standardized test. Second, by their very nature, standardized tests employ norm-referenced measurement. In other words, a child's performance is judged relative to normative data gotten from large samples of normal children. While NRM may be of great use in determining initial placement of a child in a special class in terms of his deviation from the standardized data of normal children, such comparisons are of little use in gauging the actual progress of a particular child relative to his potential. Third, NRM, or standardized testing, does not readily lend itself to measuring change in a valid manner when several measurements on the same material or module are needed; in other words, a child becomes attuned to the questions on the test itself after receiving more than one administration of the instrument. For these reasons, NRRC/P will not only use standardized tests in the usual screening of children and in classical, global program evaluation, but will emphasize a much more appropriate measurement system known as CRM.

For any given instructional module, criterion-referenced tests will be used. Such tests will be known as monitors. The basic advantage of CRM is that it gauges the progress of a pupil relative to his own potential in terms of predetermined goals of performance levels. Thus, the inappropriate comparisons that would occur in pitting an exceptional child's progress against that of normal children, that is, NRM, are avoided completely with CRM. For these reasons alone, monitors or tests constructed and interpreted in the criterion-referenced sense are ideally suited to measuring change in children as the result of highly individualized educational prescriptions. The particular objectives of any given instructional module are reflected in the test items of the monitor for that module. The test items are constructed in accord with the best measurement theory available. Both the teachers who use the monitors and modules and the measurement specialists who help build them are involved in test item selection and construction. Because achievement or performance should be measured only relative to the child's own initial baseline on that module, both pre- and post-monitors will be used for a given module. Further, if a child does not achieve on the post-monitor the degree of attainment that his potential and initial level on the pre-monitor suggest, then he will have to be recycled through the module in question with different supplementary, modular material. Again, however, the only way of measuring how successful the remediation was is to give the child a different but similarly appropriate post-monitor. Thus, several equivalent forms of monitors must be constructed for any given module. The basic functioning of the CRM system is represented in Figure 3.

NRRC/P OPERATIONAL CRM MACHINERY

When one begins to delve into the details of such a criterion-referenced monitoring system, one of the first questions to be answered is how both the monitors and instructional modules are constructed, since these two items are

FIGURE 3.

GENERAL TYPES OF DEVICES FOR INDIVIDU
ACHIEVEMENT MONITORING AND GLOBAL PROG
EVALUATION.

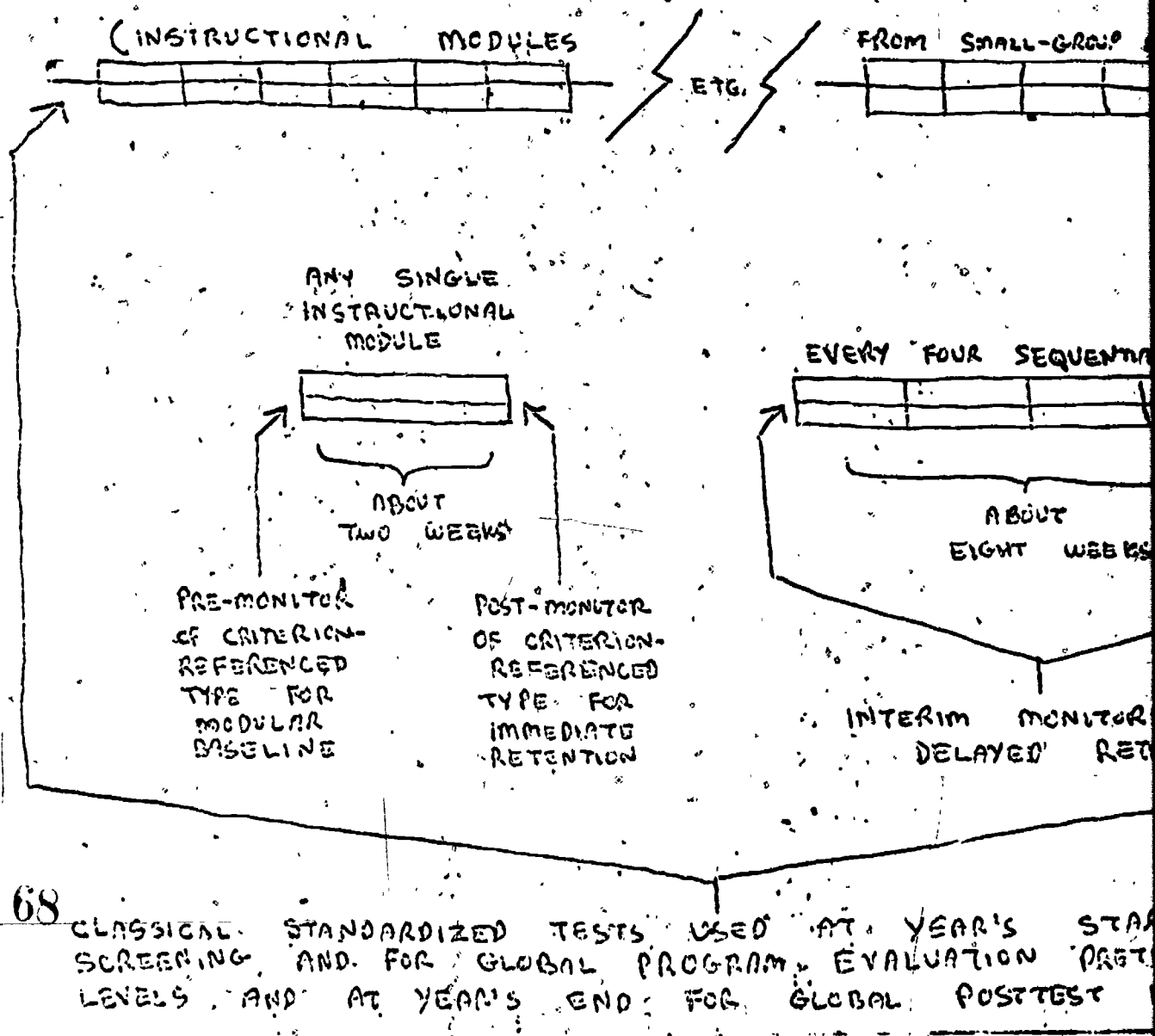
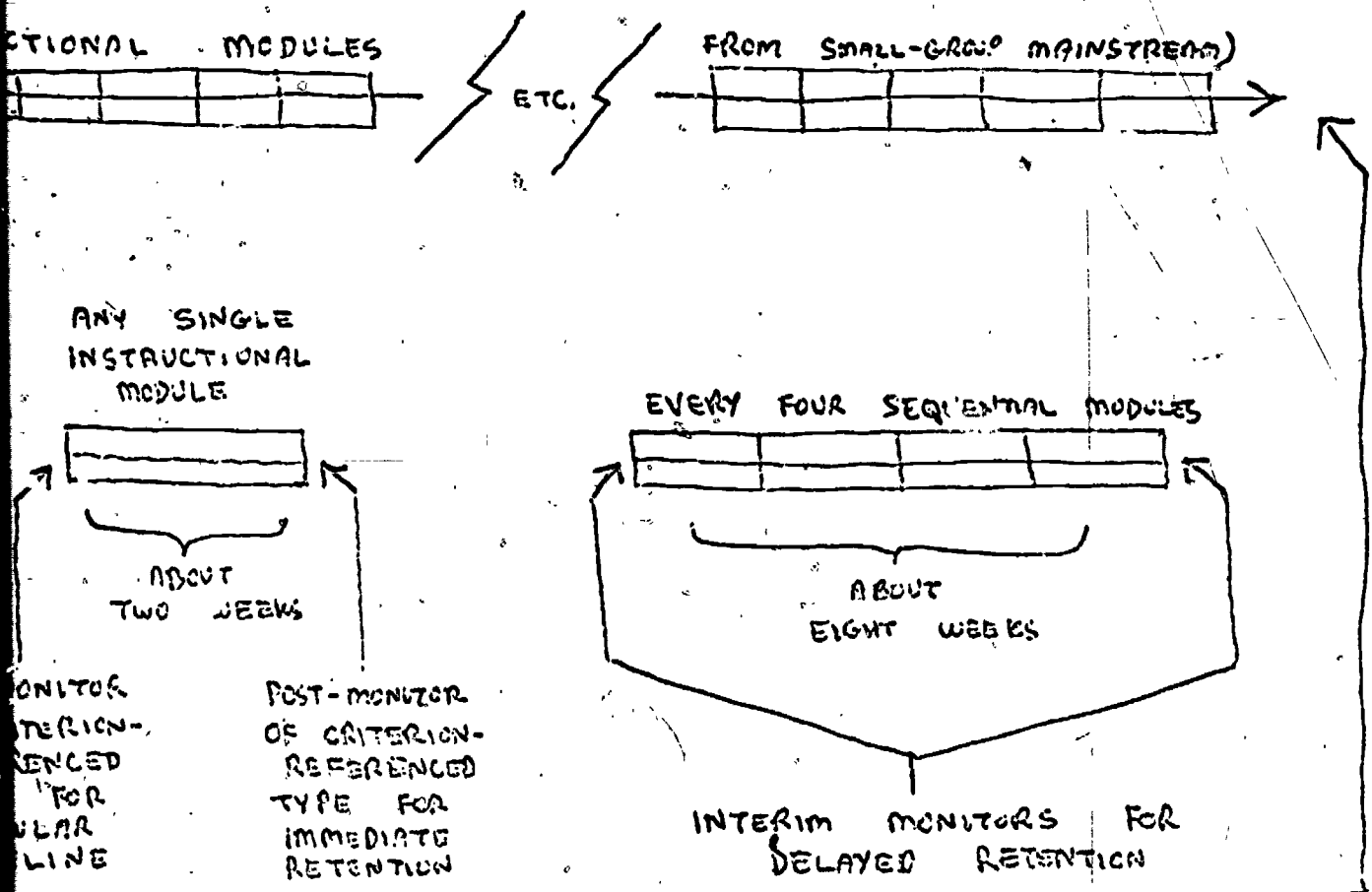


FIGURE 3

...ERAL TYPES OF DEVICES FOR INDIVIDUAL ...IEVEMENT MONITORING AND GLOBAL PROGRAM EVALUATION



STANDARDIZED TESTS USED AT YEAR'S START FOR ... AND FOR GLOBAL PROGRAM EVALUATION PRETEST ...T YEAR'S END FOR GLOBAL POSTTEST LEVELS

the heart of the program. One logical way to handle this task would be first to examine the range of abilities and deficits in the population of learning disabled pupils being served. With such a survey completed, the educational programer/prescriber can project roughly just what total range of subject content areas can be expected to be mastered throughout the year. Then, following this line of reasoning, during the summer before the start of the program, a task force of teachers, administrators, evaluators, and other specialists would work feverishly to complete a sufficient number of sequentially related instructional modules that would take care of the projected range of all pupils, along with corresponding sets of monitors. However, NRRC/P will not elect to go this route. First, the job of trying to anticipate how far each child will go throughout the year and then building enough modules to cover this wide range, is far too complicated to accomplish with adequate quality simply during the summer. Second, devising all the modules and monitors ahead of time tends to lock staff members into a "canned" set of programs that will tend to stifle in-process improvements dictated by spontaneous problems that always seem to arise. Thus, a more flexible monitor-module production system is needed for NRRC/P.

Before describing how NRRC/P plans to devise the modules and monitors, the reader should be aware of how pupils relate to each other as they move from one instructional module to the next. First, let the reader assume that all pupils in a given class are able to be handled adequately by the regularly assigned teacher in the modified small-group setting. Nonetheless, it must be borne in mind that each child is treated as an individual and is allowed to move at his own rate through whatever module appears to be appropriate to him at his stages of developmental readiness, existing knowledge, and ability. In other words, at any given point in time, each child probably will be working on a different module. However, eventually every child will pass through

some of the same modules, since they are devised in a sequential task framework. Also, since a given module will always have its corresponding monitors used with any child that passes through the module, eventually each module will have comparable data obtained from every child in the class. Of course, with some of the easier modules and some of the more advanced modules, only a few children will ever work their way through them and resulting data will be sketchy for the class as a whole at these points in the instructional sequence. But then this is the nature of individualized instruction! The sequence of steps that any child who is functioning adequately within the regularly assigned classroom setting would go through is given in Figure 4. Before returning to a discussion of how monitors and modules are constructed, let the reader consider the occasions when a child becomes so embroiled in his learning difficulties that he must be referred to the resource teacher for several days or even longer.

Highly specialized, individualized help must be provided to any child who runs into severe educational problems. In general, a resource teacher-consultant will be called in. There are at least two ways in which this can occur. First, an itinerant resource teacher will be brought into the child's regular class to work with him in that setting. Second, the child will be taken out of his regularly assigned class and sent to a special resource teacher room for a certain amount of time each day. Regardless of the particular method selected, the relationships between the resource teacher consultation and the regularly assigned small-group instruction (individualized "mainstream") are represented in Figure 5. One can see how the decision is made as to when the child returns to the small-group instructional setting.

Next, the matter of how the monitors and modules might best be constructed needs to be considered. As in the first method described above but rejected by NRRC/P, a survey of the range of abilities and weaknesses of each child

FIGURE 4

SMALL-GROUP MAINSTREAM OPERATIONAL SEQUENCE

INDIVIDUAL ACHIEVEMENT MONITORING SYSTEM

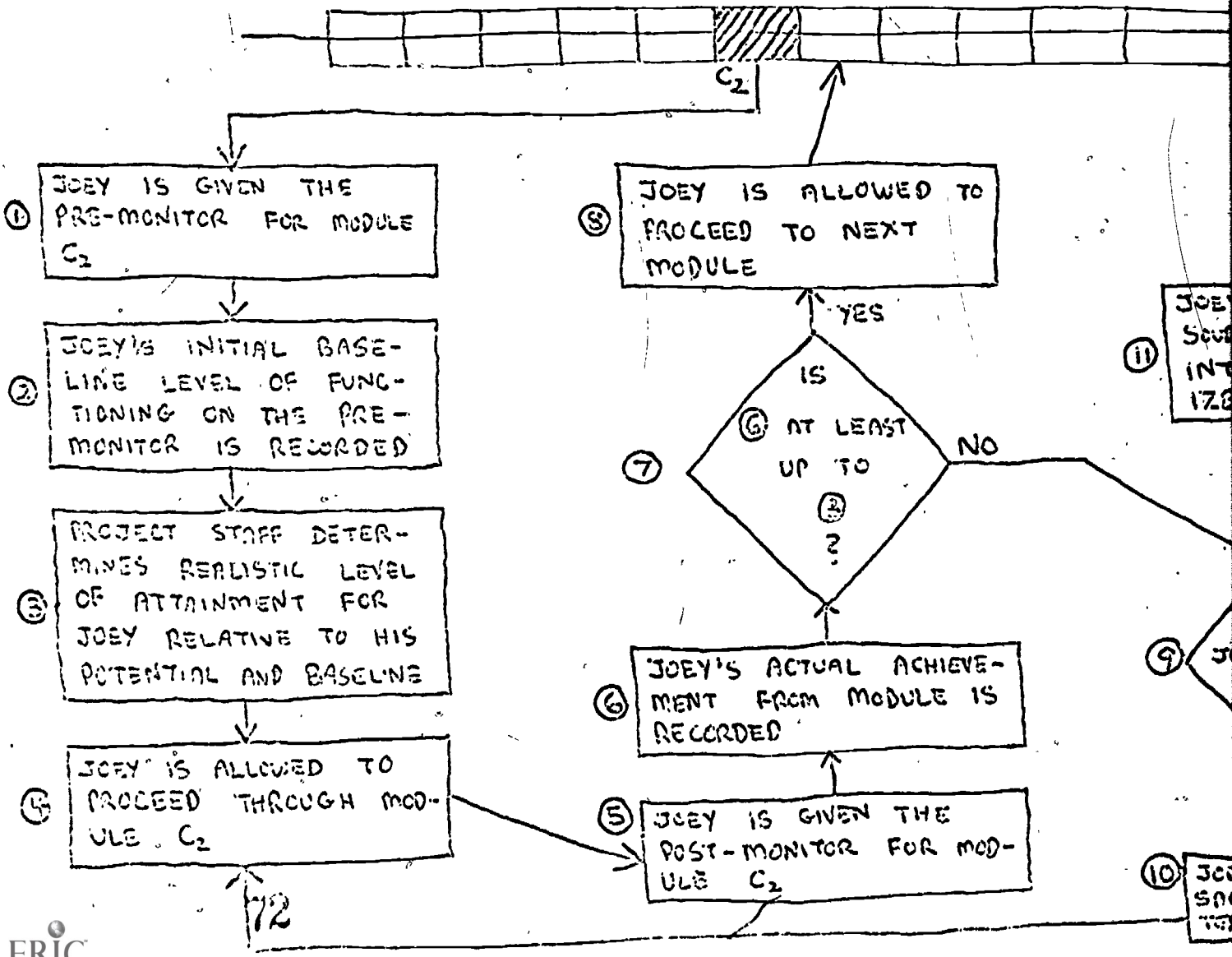


FIGURE 4

L-GROUP MAINSTREAM OPERATIONAL SEQUENCE OF
IDUAL ACHIEVEMENT MONITORING SYSTEM

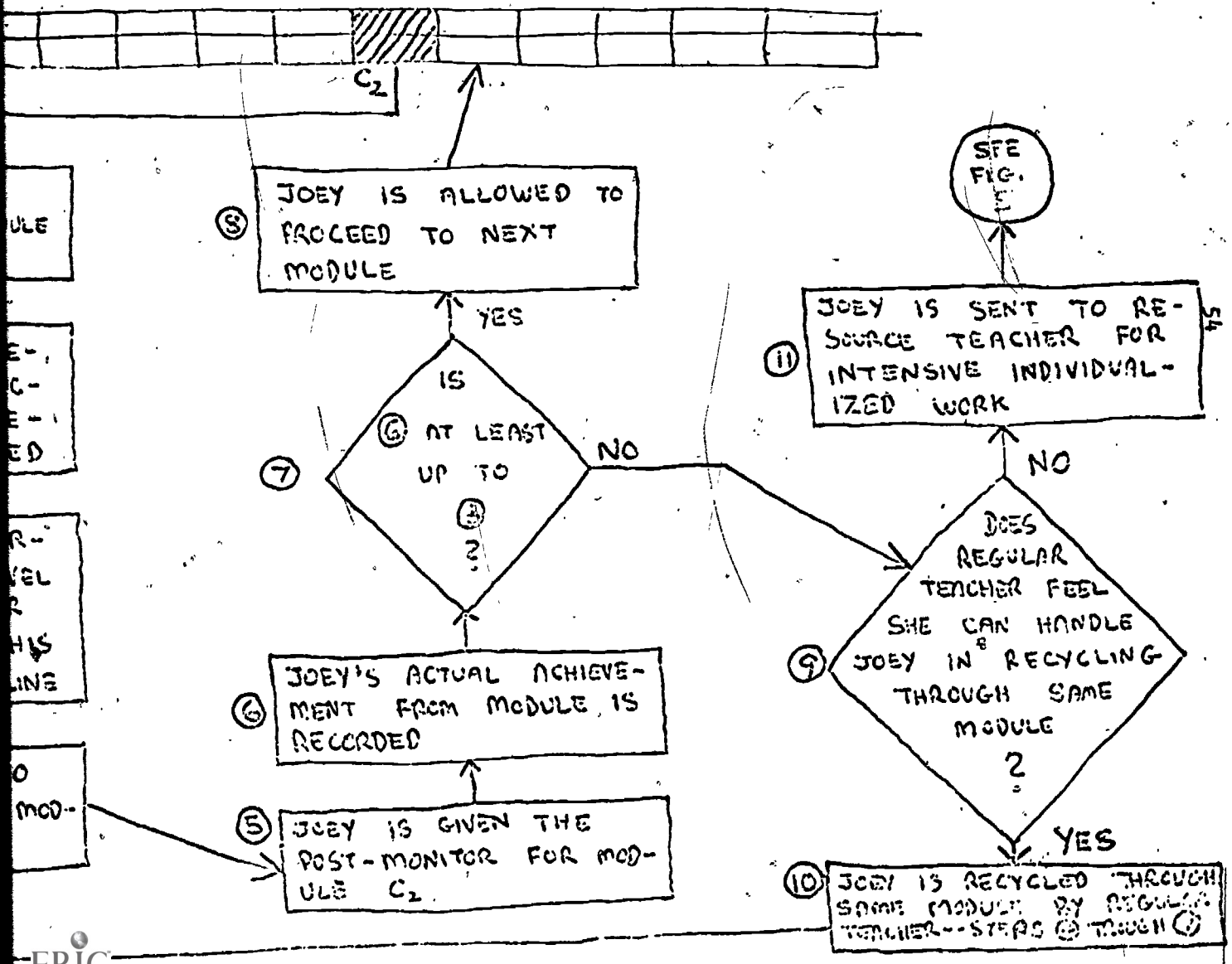


FIGURE 5

RESOURCE TEACHER OPERATIONAL SEQUENCE OF INDIVIDUAL ACHIEVEMENT MONITORING SYSTEM

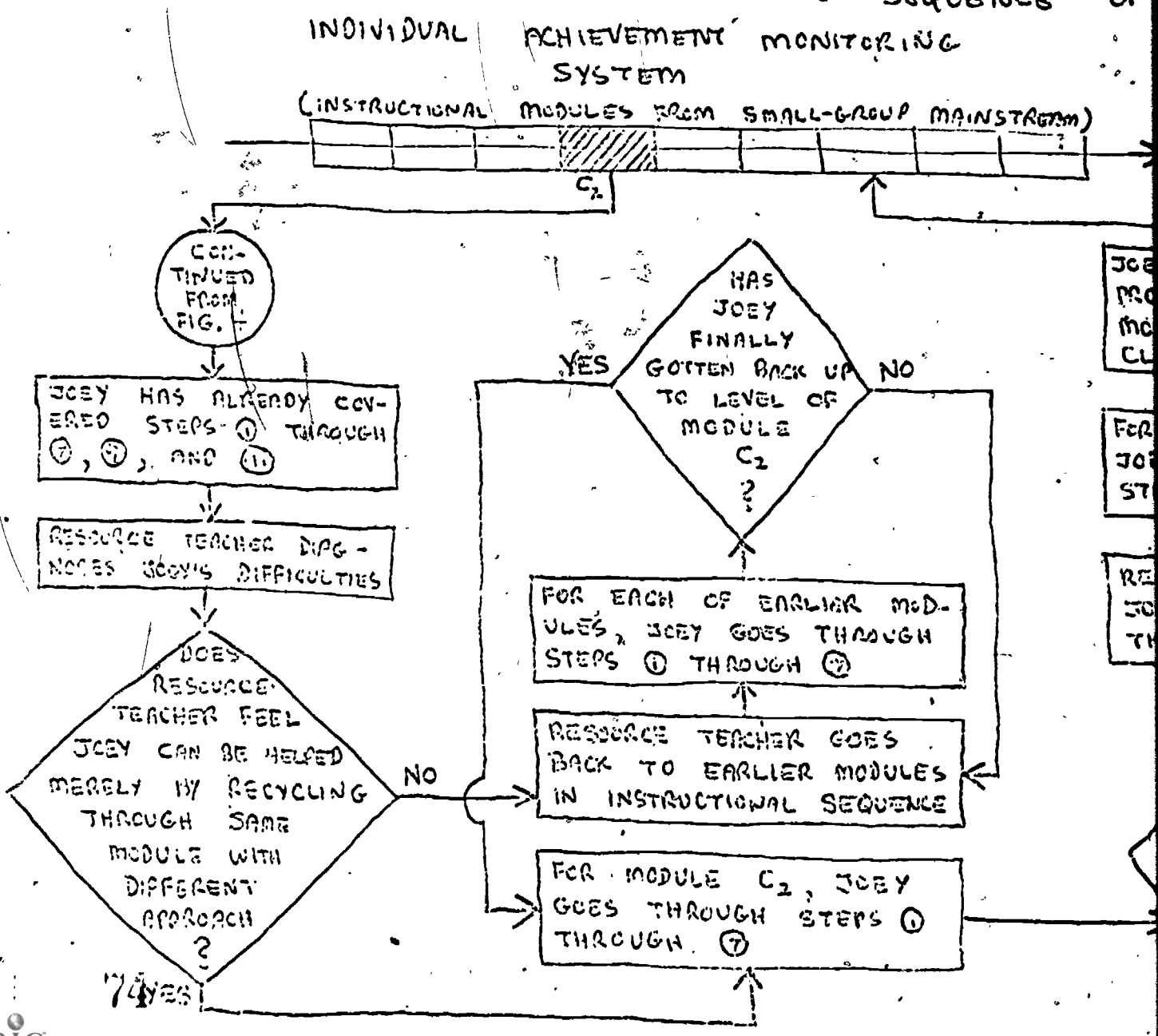
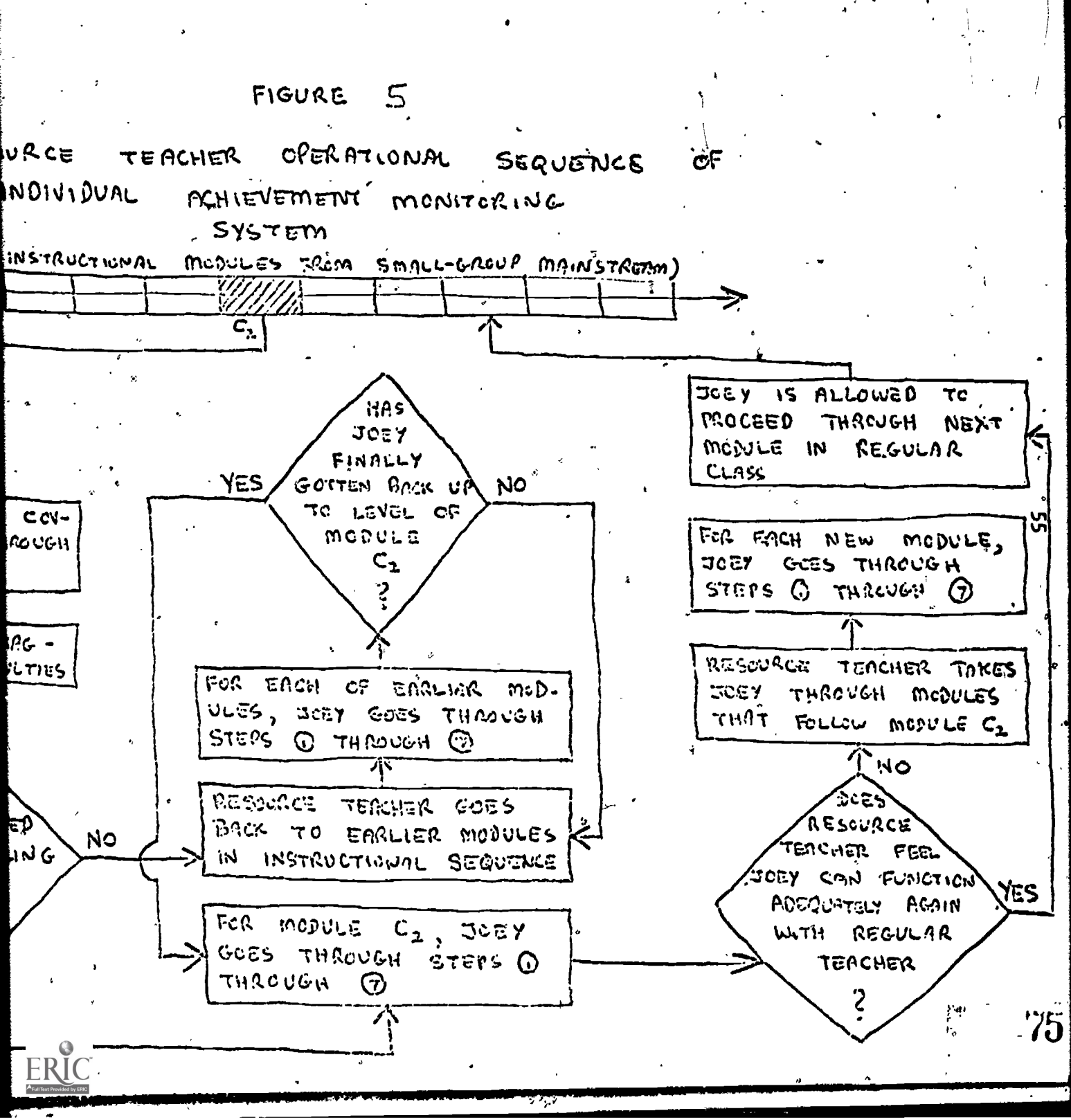


FIGURE 5

URCE TEACHER OPERATIONAL SEQUENCE OF
NDIVIDUAL ACHIEVEMENT MONITORING
SYSTEM



involved in NRRC/P experimental classes will be undertaken. Looking at the initial range of probable starting points of each child in the instructional sequence, only those modules and corresponding monitors that will be needed at the start of the school year to accommodate every child's starting point, will be constructed during the summer in special work sessions. However, beyond these initially required modules and monitors, none will be constructed until shortly before the need arises during the regular academic year.

Weekly meetings of teachers, measurement specialists, and programming personnel will be held throughout the year to discuss the immediate modular needs of each teacher for each pupil. Thus, the system of constructing modules and monitors is kept completely flexible. The modular needs of each teacher are anticipated only a week or two in advance of actual usage. Also, mutual feedback about in-process problems, staff criticisms, etc., is accomplished at these weekly meetings. The manner in which modular needs are anticipated and met is simply a run-down from each teacher of where she feels each child in her class will be in the instructional sequence in the next week or two. Also, whenever a child has been entrusted to the resource teacher situation, that specialized programming consultant must make known his modular needs, too. In this way, all children are covered at all times.

At any given point, once the modular needs have been assessed for all children, such feedback is handed over to the programming specialists. For example, NRRC/P intends to concentrate only on reading and arithmetic during its first operational year. Thus, programming specialists in these two fields will have primary responsibility for devising the instructional modules. Before modules are put into practice, teachers will have a chance to recommend changes wherever they see difficulties.

As part of the instructional module construction process, monitors corresponding to each module must be devised concurrently. Briefly, the

monitors begin to approach reality when the specific objectives (stated in measurable terms) of the module in question are agreed upon. A large test item pool for all objectives is constructed, with individual test items coming from already existing sources or being made on the spot. At least three parallel forms of a monitor are constructed for a given module. The first parallel form of the monitor serves as a pre-test, the second as an immediate post-test, and the third as a second post-test (if the child has to be recycled through the same module again because of inadequate criterion performance on the first post-test). The three parallel forms are constructed by randomly assigning test items for a given objective throughout all three forms. This process is repeated for each objective in the module.

Another logical question one might ask about specific operational procedures concerns personnel for carrying out the monitoring process. While it will be the responsibility of administrative staff to provide the teachers with raw working materials -- modules, monitors, instructional materials, etc. -- teachers themselves will be required to administer each monitor to each child whenever the appropriate time arises. Only the teacher will be able to coordinate this activity most efficiently. Further, the teacher will be required to grade the monitors herself and to record all data on specially devised recording sheets. In this way, the teacher becomes intimately involved in diagnostic teaching and provides herself with immediate feedback on how each child is doing with the module he is currently involved with. Indeed, the big advantage is that the teacher is forced to look at just how each child is learning; in other words, accountability (see Proger, in press) becomes reality. On each teacher's recording sheet, data will be kept on monitor scores (in relation to predetermined criterion levels of success custom-made for each child), time needed to go through the module, open-end comments of the teacher, etc.

Every week an administrative supervisor will collect a carbon copy of the data recording sheet of each teacher. This carbon copy is returned to the evaluation department for processing and analyzing. All data will be entered on computer cards according to a predetermined format. Eventually, after a large number of children go through the same module with the same corresponding monitors, evaluation personnel will be able to draw some generalizations about how certain types of children learn the modular subject-matter material in question. Reports will be generated and disseminated at local, state, and national levels.

SUMMARY

An individual achievement monitoring system for special education has been described. The criterion-referenced nature of this monitoring system has been explained, in distinction to the usual norm-referenced measurement procedures of standardized testing. The details of this CRM system as projected for use in the National Regional Resources Center of Pennsylvania project have been outlined.

CHAPTER VII

"MACHINERY" FOR IMPLEMENTATION OF THE FORMAL PROGRAM EVALUATION SYSTEM FOR SPECIAL EDUCATION AT THE STATE LEVEL: PERSONNEL AND DATA-BANKING ACTIVITIES

STATE IMPLEMENTATION MACHINERY

The previous section has dealt with regional implementation of the formal program evaluation model. I personally feel that local consultation agencies are definitely crucial to any statewide system. However, regardless of how local special education personnel obtain expert consultation on how to carry out their particular program evaluation activities, all data collected must be assimilated into the state Bureau of Special Education, analyzed, interpreted, and used for policy decisions wherever appropriate. In this section, the implementation machinery at the state level will be considered.

First, the state Bureau of Special Education would need some specialized evaluation and measurement personnel. Ideally, such people would be well-versed in evaluation and measurement methodology, statistical analysis, and data processing and computer programming. There are two primary sources for obtaining such people. First, the United States Office of Education has turned out hundreds of such specialists at the doctoral level through the Educational Research Fellowship Training Program. Second, several colleges and universities have Master's programs in educational research. There should be no difficulty in getting hold of qualified personnel.

The next question to be resolved concerns where such program evaluation specialists would be housed and from what budgets they would be paid. There is no doubt that major planning, policy, and operational decisions should be made by evaluation specialists who have had extensive training and experience at the doctoral level. Such personnel need not necessarily be paid by or housed within the Bureau itself. Evaluation specialists of doctoral calibre can be gotten on a consultant basis (paid or non-paid) from universities or special federal projects. In fact, no doubt special federal projects could be initiated whose sole function would be to field test the feasibility of such a program evaluation model; the Director of the Bureau could be made Principal Investigator so as to maintain Bureau control over the evaluation activities.

Once expert measurement personnel at the doctoral level have been obtained on a part-time consultant basis or full-time basis, more routine operational details could be accomplished by measurement specialists at the Master's level. Again, salaries and facilities could be handled directly out of the Bureau, or special agencies outside the state Department of Education could be funded (perhaps federally) to handle such evaluation tasks of a day-to-day nature with regard to data processing and analysis.

After one has considered the question of personnel in depth, he must next deal with facilities for storing and analyzing data and information obtained from separate evaluation projects. Such computer facilities already exist within the state Department of Education and a large number of county or intermediate unit operations. Cooperative arrangements could be explored at the state and regional levels. Possibilities in colleges and universities would also be considered. Keypunching facilities devoted exclusively to a statewide program evaluation system in special education are a necessity. In collecting the data for storage and analysis, a feasible, standardized

format for inputing the data must be devised. All such operational details could be handled by measurement specialists and other consultants of whom the Bureau of Special Education would want to avail themselves.

In terms of both personnel and facilities for processing and analyzing, the obvious suggestion would be to use existing arrangements, wherever satisfactory, within the state Department of Education and state-affiliated organizations and projects. Nonetheless, wherever currently available resources would clearly not be able to handle the tasks of a formal program evaluation system, then specialized personnel and facilities must be obtained that would be devoted solely to special education purposes.

This chapter will be concluded by devoting some detailed comments to the concepts of data-banking. To give the reader some working ideas of just what data-banking activities consist, a review of the literature is presented.

One of the key features of the NRRC/P research program will be its emphasis on CRM, although not to the complete exclusion of standardized or NRM testing. In order to answer the types of questions posed at the start of this paper, it is imperative that a large amount of information be stored so that it can later be retrieved for various types of pooling operations via different analytical strategies. The data bank (DABA) concept is a vehicle for such activities. A brief review of the technical literature in this field will be helpful for those interested in the functioning of NRRC/P.

The need for a DABA in a large project engaged in massive testing programs is apparent and yet feasibility studies with the DABA idea are scarce. Austin (1970, p.9) claims: "Those of us who are concerned with the processing of large scale testing programs have, in the past few years, made considerable progress in the area of high-speed test scoring. ... In the area of record-keeping, or data banking, we have done little."

Since the establishment of a data bank (DABA) is one of the primary

enabling objectives of NRRC/P in working toward the questions posed at the start of this paper, a few comments on operational difficulties are in order. Fascione and Penry (1970) describe the experiences of the School District of Philadelphia in trying to generate a data bank. They point out that not many educators, let alone other types of technicians, are familiar with just what DABA implies. Further, they warn against getting up in the sometimes grandiose ideas of systems analysts, computer workers, etc. Fascione and Perry (1970) suggest. "Primary responsibility for system design and implementation both should be placed somewhere in the organizational structure other than in the data processing area. This crucial initial step helps to retain the project's focus on the human aspects of the problem. Another way of describing the benefits of this approach is to say that it helps prevent the tail from wagging the dog, which results from the data processing technologists' natural inclinations to (1) have the latest and most sophisticated equipment, (2) justify the computer's presence by utilizing all its capacity, (3) set the actual goals for the system rather than have them set by administrators." In line with this philosophy, Fascione and Penry recommend that DABA managers set out to produce immediate benefits for those to be served, rather than harping on what tremendous things will happen with long-range goals. Some immediate benefits that resulted in the School District of Philadelphia were: (a) compilation of student attendance and background lists for administrators, (b) capacity to conduct longitudinal studies of certain children, (c) capacity to draw more valid and representative random samples of certain types of students for ongoing evaluation studies, and (d) keeping track much more efficiently of standardized testing results. Fascione and Penry describe their DABA system as being based mainly upon cards rather than tapes. Each child is kept on a card, with such things as background characteristics as name, birth date,

sex, ID number, address, home telephone number, school grade, room assignment, etc. An up-to-date punched card deck is maintained within each school building. Every two months these card decks are re-pocessed in entirety to give up-dated lists of students.

The CAM Project at the Univers' :y of Massachusetts also maintains what is, in effect, a DABA (cf. Gorth, Grayson, Popejoy, and Strowd, 1969). With a CRM system such as CAM, a ' ge amount of performance data is obtained with respect to individual test items, groups of test items relating to one behavioral objective, groups of behavioral objectives relating to one large program objective, and -- turning along a different dimension of data generation -- data on individual students, groups of students, etc. The comparisons are almost endless. Thus, the needs for a highly efficient DABA are evident.

Perhaps the classic example of the sophisticated DABA is that associated with Project TALENT (Flanagan, Cooley, Shaycroft, Hall, VanWormer, Wingersky, and Holdeman, 1965). "Although the term 'data bank' is sometimes used to refer to any accumulation of data, it is important to recognize that some accumulations will be more useful than others. It seems preferable to reserve the term 'data bank' for data collected with some over-all basic design and for which research uses were originally considered. This does not necessarily mean that the data must have been collected solely for research purposes, but it does mean that no sound research principles were violated in the data-collection process. (p. 1)"

Flanagan et al. (1965) list seven features that they consider essential to a meaningful data bank: (a) the data gathered must relate to a population of students that has been previously defined in a deliberate and careful manner with randomization present, rather than inadvertently defined populations; (b) As many variables as possible should be tapped; (c) If

possible, a large number of variables should be measured on a large sample; (d) data in the bank should be easily accessible; (e) all data collected should be comparable with respect to type of instrument, time of administration, conditions of measurement, etc.; (f) data should be organized within the bank so that complex relationships can be derived by computer; and (g) data recorded at different points in time should be interrelated for the same students, and any factors that may have affected such relationships should also be able to be tied in. Flanagan et al. state: "The administration of the Project TALENT tests to nearly half a million students in over 1,300 schools constituted the first of several phases of data collection. More than 2,000 items of information per student and 1,000 items per school were collected. Some of these have been summarized in the form of test scores and others have been transferred directly to magnetic tape, currently stored at the Computation and Data Processing Center of the University of Pittsburgh. A series of follow-up studies has been planned for one, five, ten, and twenty years after each of the (four) classes in the sample graduates from higher school. (p. 4)" Thus, long-range career patterns will be able to be related to original patterns of education, as well as a host of other variables. This was one of the main objectives of Project TALENT.

The NRRC/P DABA will hardly be as extensive as Project TALENT's DABA, but the concepts of operation will be highly similar. The ideas of collecting periodic information on what is happening throughout the remediation process used with the student and trying to relate such data to different strategies of remediation, as well as background variables on the student, will be a primary goal.

One of the most exhaustive studies of the educational DABA idea was the series of reports contained in Carroll et al. (1965). A series of conferences held by the Harvard Graduate School of Education debated issues connected with

the DABA concept. In one report, Benjamin Bloom (pp. 30-37) discussed some problems: (a) the originators of a particular DABA determine at the outset what types of information are most important; (b) whether the DABA should act as a service center or a research center; (c) the possible conflicts between individual research efforts and team research projects; (d) the possible invasion of privacy; (e) whether DABA's evolve over time with improvements or maintain their original structure. As an example of an example of an existing DABA, Bloom mentioned the International Educational Achievement Study Test results in mathematics were collected for 200,000 students of ages 13 to 17 or 18 from the United States, eight European countries, Israel, Japan, and Australia.

In the DABA report of Carroll et al. (1965), a conference of school superintendents resulted in recommendations of the types of questions they would like answered. Two examples (p. 45) are (a) "What is the relationship between subjects or courses of study pursued in high school and the occupation the student enters after graduation?" and (b) "What pre-school experiences best prepare a child for school experience, especially with regard to reading and motivation to learning?" However, the superintendents pointed out that a DABA has inherent limitations because "for practically every question considered to be of great importance, the data to answer the question were virtually inaccessible. ...inaccessibility does not mean that data do not exist, but rather that the effort required to retrieve them or rearrange them manually would be so great that the data are, for all practical purposes, not at all accessible (p. 46)"

Another example of a functioning DABA given by Carroll et al. (1964) is the New England Education Data Systems (NEEDS). The report states: "the realities of running a school -- such things as production of schedules, report cards, class lists, attendance records -- because of their immediacy,

require attention and time. NEEDS seeks to provide ways to reduce the time and attention taken by these clerical tasks, thereby releasing the administrator and his staff for more important, creative work such as assessment and reorganization of the curriculum (p. 73)." NEEDS includes nine communities in Massachusetts, two in Connecticut, two in Vermont, one in Rhode Island, and one in New Hampshire. The four divisions of NEEDS are (a) data processing services, (b) operations research and development, (c) in-service training, and (d) basic research and formal instruction. The basic services offered are (a) file creation and maintenance, (b) scheduling support, (c) mark reporting, (d) automated attendance, and (e) test scoring and analysis.

A second major illustration of a functioning DABA found in Carroll et al. (1965) is the Iowa Educational Information Center (IEIC), sponsored by the College of Education at the University of Iowa and the State Department of Public Instruction. The data banking and data processing activities of IEIC are similar to those of NEEDS.

The report of Carroll et al. (1965, p. 20) recommended that at least three types of data be considered for any DABA: (a) "demographic data (age, sex, socio-economic status of parents, and other data which are essentially sociological)," (b) "descriptive data (class size, pupil-teacher ratio, and other summary statistical data which describe characteristics of the school or the student population, personnel, etc.)," and (c) "evaluative data (tests, student grades, and other data for evaluation of student progress, teacher success, curricular validity, etc.)." The report concludes with an extensive bibliography of DABA literature.

Miami (Dade County, 1967) has taken the lead in a statewide DABA operation. The system will (a) provide teachers with periodic background reports on students, (b) help curriculum planners evaluate particular programs, (c) establish mutual feedback between the schools and colleges, and (d) provide

guidance counselors with student information reports. Using Miami as a pivot point, four different counties in Florida tried out different techniques associated with a DABA system to test the feasibility of a statewide DABA operation.

Other examples of the DABA concept are readily found and will not be detailed here. The value of such systems, when and if they become sophisticated enough, is that any type of remediation used with a student can be evaluated in terms of the effects it had on the student relative to other approaches. (cf. Grossman and Howe, 1966; McComb, Miss., 1967; St. Louis Park, Minn., 1967; Sacramento, Calif., 1966; Edina, Minn., 1966; Mount Clemens, Mich., 1967; Davenport, Iowa, 1966; Lincoln, Nebr., 1967; Buffalo, N.Y., 1966; Eugene, Oreg., 1966).

REFERENCES

Alkin, M.C., Glinski, R., and Wininger, R.: Preliminary Analysis of Data for a Secondary School Input-Output Model. Los Angeles: Center for the Study of Evaluation, UCLA Graduate School of Education, 1969 (CSE Report No. 42).

Allen, Dwight W. Significant Differences: On the Social Insignificance of Statistical Significance -- A plea for New Strategies of Evaluation. Educational Researcher: Official Newsletter of the American Educational Research Association, 1969, 20 (8), 3-4.

Allen, Dwight W. Stimulating Change in Instructional Systems through New Evaluation Techniques. Paper presented at the Annual Meeting of the American Educational Research Association, Minneapolis, Minn., March, 1970.

Allen, Dwight W., O'Reilly, Robert P., and Gorth, William P. An Introduction to Longitudinal Testing Using Item Sampling Techniques. Paper presented at the Annual Meeting of American Educational Research Association, Minneapolis, Minn., March, 1970.

Ammentorp, W., Daley, M. F., and Evans, D. N.: Prerequisites for systems analysis: analytic and management demands for a new approach to educational administration. Educational Technology, 1969, 9(8), 44-47.

Annas, P. A., and Dowd, R. A. (Eds.): Guide to Assessment and Evaluation Procedures: The New England Educational Assessment Project. 1966 (Available through ERIC Document Reporduction Service, No. ED012067).

Atkinson, G.: Evaluation of educational programs: an exploration. In: Strevell, W. H. (Ed.). Rationale of Education Evaluation. Pearland, Texas: Interdisciplinary Committee on Education Evaluation, Gulf Schools Supplementary Education Center, 1967. pp. 1-9.

Austin, Gilbert, State Directors Discussions: Test Scoring, Reporting and Data Banking. NCME Measurement News: Official Newsletter of the National Council on Measurement in Education, 1970, 13 (3), 9.

Bayuk, Robert J., Jr., Proger, Barton B., and Mann, Lester, Organization of Meaningful Verbal Material. Psychology in the Schools 1970, 7, 365-369.

Bloom, B. S. (Ed.) Taxonomy of Educational Objectives: Handbook 1: Cognitive Domain. New York: David McKay, 1956.

Brooks, C. N.: Training system evaluation using mathematical models. Educational Technology, 1969, 9(6), 54-61.

Buffalo, New York, Board of Cooperative Educational Services, Erie County Regional Educational Data Processing and Information System. Buffalo, N. Y.: Author, 1966. (available through ERIC as document #ES-000-414).

CAPS: A bibliography for national assessment. CAPS Capsule, 1970, 3(2), 13.

Carroll, J. B. A Model of School Learning. Teachers College Record, 1963, 64, 723-33.

Carroll, John B., et al. Planning and Utilization of a Regional Data Bank for Educational Research Purposes: Final Report, Cambridge, Mass.: Harvard University, Laboratory for Research in Instruction, 1965 (available through ERIC as document #ED-003-480).

Carter, L. F.: The systems approach to education: mystique and reality. Educational Technology, 1969, 9(4), 22-31.

Center for Instructional Research and Curriculum Evaluation (CIRCE) and Cooperative Educational Research Laboratory, Inc. (CERLI): Information supplement #5: evaluation kit: tools and techniques. Educational Product Report, 1969, 2, 16pp.

Cox, Richard C., and Sterrett, Barbara G. A Model for Increasing the Meaning of Standardized Test Scores. Journal of Educational Measurement, 1970, 7, 227-228.

COX, Richard C. Evaluative Aspects of Criterion-Referenced Measures. Paper presented at Annual Meeting of the American Educational Research Association, Minneapolis, Minn., March, 1970.

Cronbach, L. J.: Course Improvement through evaluation. Teachers College Record, 1963, 67, 672-681.

Dade County Board of Public Instruction. Improved Educational Services and Practices through Utilization of Electronic Records. Miami, Fla.: Author, 1967. (available through ERIC as document #ES-001-857).

Davenport, Iowa, Scott County Board of Education. Area IX Total Information System. Davenport, Iowa: Author, 1966. (available through ERIC as document #ES-000-231).

Dyer, H. S.: An educational researcher's view of systems analysis. Educational Technology, 1969, 9(9), 83-85.

Dyer, H. S., and Solomon, R. J.: Statewide assessment: its future and potential for educational reform. Paper presented at the Annual Meeting of the American Educational Research Association, Minneapolis, Minn., March 1970.

Ebel, R. L.: Assessing national assessment. CAPS Capsule, 1970, 3(2), 10, 12.

Edina, Minn., Suburban School Service Joint Board. Coordinated Data Processing Service and Facility. Edina, Minn.: Author, 1966. (available through ERIC as document #ES-000-251).

Emrick, John A., and Adams, E. N. An Evaluation Model for Individualized Instruction. Paper presented at the Annual Meeting of the American Educational Research Association, Minneapolis, Minn., March, 1970.

EPIC: Generalized scheme for evaluations of innovations. The EPIC Forum, March-April, 1968, No. 2, 1-2 (Tucson, Arizona: Educational Programs for Innovative Curriculums).

Erickson, D.: Educational accountability. IMC Reports: New York State Network, Special Education Instructional Materials Centers, 1970,2(3), 1.

Eugene, Oregon, Lane County Board of Education. Oregon Total Information System (OTIS). Eugene, Oreg.: Author, 1966 (available through ERIC as document #ES-000-534).

Fascione, Daniel R., and Penry, Edward B. Establishing a Pupil Data Bank: Conflict Between Theory and Practice. Paper presented at Annual Meeting of the American Educational Research Association, Minneapolis, Minn., March 4, 1970.

Findley, W. G.: To know or not to know: that seems to be the question. CAPS Capsule, 1970, 3 (2), 9-10.

Finn, J. D.: Institutionalization of evaluation. Educational Technology, 1969, 9(12), 14-23.

Flanagan, John C., Cooley, William W., Shaycroft, Marion F., Hall, Charles E., Van Wormer, Joh, Wingersky, Bary G., and Holdeman, Richard W. The Project TALENT Data Bank: A National Data Bank for Research in Education and the Behavioral Sciences. Pittsburgh, Pa.: Project TALENT, University of Pittsburgh and American Institutes for Research, 1965.

Freedman, S. A., and Swanson, J. R.: A model for planning education: needs assessment/programming/implementation/decision making. Non-Technical Paper, April 22, 1969, No. 3 (Tallahassee, Florida: Division of Research, State Department of Education).

Gorth, William P. A Synergistic Relation Between Teachers and Evaluation. Paper presented at Annual Meeting of the American Educational Research Association, Minneapolis, Minn., March, 1970.

Gorth, W. P., Grayson, A., and Lindeman, D. A Computer Program to Evaluate Item Performance by Internal and External Criteria in a Longitudinal

Testing Program: Using Item Sampling. Educational and Psychological Measurement, 1969, 29, 181-183.

Gorth, W. P., Grayson, A., Popejoy, L., and Strowd, T. A Tape-Based Data Bank for Educational Research or Instructional Testing Using Longitudinal Item Sampling. Educational and Psychological Measurement, 1969, 29, 175-177.

Gorth, W. P., Grayson, A., and Strowd, T. A Computer Program to Tabulate and Plot Achievement Profiles of Longitudinal Achievement Testing Using Item Sampling. Educational and Psychological Measurement, 1969, 29, 179-180.

Griessman, B. E.: An approach to evaluating comprehensive social projects. Educational Technology, 1969, 9(2), 16-19.

Grobman, Hulda: Evaluating Activities of Curriculum Projects: A Starting Point. Chicago: Rand McNally, 1968 (AERA Monograph Series on Curriculum Evaluation, No.2).

Groff, P.: Super test, or how to stop worrying and enjoy the national assessment. Educational Researcher, 1970, 21 (March), 5-8.

Grossman, Alvin, and Howe, Robert L. Research and Development in Data Processing for Pupil Personnel and Curricular Services. Sacramento, Calif.: California State Department of Education, 1966. (available through ERIC as document #ED-010-618).

Guba, E. G.: The failure of educational evaluation. Educational Technology, 1969, 9(5), 29-38.

Guilford, J. P. The Nature of Human Intelligence. New York: McGraw-Hill, 1967.

Hammond, R. L.: Evaluation at the Local Level. Tucson, Arizona: Project EPIC, n.d.

Hammond, R. L.: Context evaluation of instruction in local school districts. Educational Technology, 1969, 9(1), 13-18.

Kari, Marion: An exampie of process evaluation. Paper presented at the Annual Meeting of the American Educational Research Association, Minneapolis, Minn., March, 1970.

Katzman, M. T.: Assessment of current systems vs. inventing new educational technologies. CAPS Capsule, 1970, 3(2), 8-9.

Katzman, M.T., and Rosen, R. S.: The science and politics of National Educational Assessment. Teachers College Record, 1970, 71, 571-586.

Kearney, C. P.: The politics of educational assessment. Paper presented at the Annual Meeting of the American Educational Research Association, Minneapolis, Minn., March, 1970.

Kennedy, John J. A Significant Difference Can Still Be Significant. Educational Researcher: Official Newsletter of the American Educational Research Association, 1970, 21 (October), 7-9.

Klein, Stephen. Evaluating Tests in Terms of the Information They Provide. Evaluation Comment (Center for the Study of Evaluation, UCLA), 1970, 2 (2), 1-6.

Kunzelman, Harold P. Data Diagnosis and Programming: Part III: Data Decisions. In John I. Arena (Ed.), Selected Papers on Learning Disabilities: Successful Programming: Many Points of View. Pittsburgh, Pa.: Association for Children with Learning Disabilities, 1969 (Fifth Annual International Conference, Boston, Mass., February 1-3, 1968), pp. 428-435.

Lincoln, Nebraska, City School District., Assistance in Decision making through Retrieval in Education. Lincoln, Nebr.: Author, 1967. (available through ERIC as document #ES-001-794).

Loadman, W. E., and Major, J. L.: Providing information for decision-makers in Michigan: compilation, analyses and reporting of assessment data. Paper presented at the Annual Meeting of the American Educational Research Association, Minneapolis, Minn.; March, 1970.

Lundin, Stephen C. A Curriculum Evaluation and Revision Based on Domain Referenced Achievement Test Systems. Paper presented at Annual Meeting of the American Educational Research Association, Minneapolis, Minn., March, 1970.

Mann, Lester, Taylor, Raymond G., Jr., Proger, Barton B., Dungan, Roy H., and Tidey, William J. The Effect of Serial Retesting on the Relative Performance of High- and Low-Test Anxious Seventy-Grade Students. Journal of Educational Measurement, 1970. 7, 97-104.

Mann, Lester, Taylor, Raymond G., Jr., Proger, Barton B., and Morrell, James E. Test Anxiety and Defensiveness Against Admission of Test Anxiety Induced by Frequent Testing. Psychological Reports, 1968, 23, 1283-1286.

Mayo, Samuel T. Mastery Learning and Mastery Testing. NCME Measurement in Education: A Series of Special Reports of the National Council on Measurement in Education, 1970, 1 (3), 1-4.

McComb, Miss., Municipal Separate School District. Using Data Processing to Evaluate and Improve Classroom Instruction in Selected Mississippi School Districts. McComb, Miss.: Author, 1967 (available through ERIC as document #ES-001-4487).

Meierhenry, W. C. (Ed.): Planning for the Evaluation of Special Education Programs. Lincoln, Neb.: Teachers College, The University of Nebraska, 1969 (Prepared under contract with U. S. Office of Education, Education for the Handicapped Branch, No. OEG-0-9-372160-3553 (032)).

Michigan Department of Education: Purposes and Procedures of the Michigan Assessment of Education. Lansing, Mich.: Michigan Department of Education, 1969 (Assessment Report No. One).

Mount Clemens, Mich., Macomb County Integrated School District. Integrated School District. Integrated Educational Information System. Mount Clemens, Mich: Author, 1967. (available through ERIC as document #ES-001-156

North Central Association of Colleges and Secondary Schools: Evaluation guide for secondary schools. North Central Association Quarterly, 1969, 43, 295-315.

O'Reilly, R. P.: State education department leadership in project and regional evaluation systems. Paper presented at the Annual Meeting of the American Educational Research Association, Minneapolis, Minn., March, 1970.

O'Reilly, Robert P., Schriber, Peter E., Gorth, William P., and Wightman, Lawrence. Draft Copy: Guide for Implementing the Comprehensive Achievement Monitoring System. The University of the State of New York, The State Education Department, Division of Research, 1969 (Ditto).

Pa. Department of Education: Phase I Findings: Educational Quality Assessment. Harrisburg, Pa.: Pa. Department of Education, 1968.

Pinsky, Paul D. A Mathematical Model for Measurement and Control of Classroom Achievement. Amherst, Mass.: Comprehensive Achievement Monitoring Project, School of Education, The University of Mass., 1970 (Working Paper WP-11, Ditto).

Pohland, P. A.: Educational ethnology and evaluation. Paper presented at the Annual Meeting of the American Educational Research Association, Minneapolis, Minn., March, 1970.

Popham, W. James, and Husek, T. R. Implications of Criterion-Referenced Measurement. Journal of Educational Measurement, 1969, 6 (1), 1-9.

Proger, B. B., Haughey, C. F., Spaans, D. N., and Proger, Helen M.: Large-scale, personalized information retrieval of psychological and educational research findings for school district decision making. Paper presented at the Seventh Annual National Information Retrieval Colloquium, Philadelphia, Pa., May 8, 1970.

Proger, Barton B., Mann, Lester, Taylor, Raymond G., Jr., and Morrell, James E. Test Anxiety and Defensiveness Experimentally Induced by Four Conditions Testing Arousal. Journal of Experimental Education, 1971 (in press).

Proger, Barton B., Taylor, Raymond G., Jr., Mann, Lester, Coulson, John M., and Bayuk, Robert J., Jr. Conceptual Pre-Structuring for Detailed Verbal Passages. The Journal of Educational Research, 1970, 64, 28-34.

Randall, R. S.: An operational application of the CIPP model for evaluation. Educational Technology, 1969, 9(7), 40-44.

Reynolds, M. C.: Approaches to evaluation. In: Knoblock, P., and Johnson, J. L. (Eds.) The Teaching-Learning Process in Educating Emotionally Disturbed Children. Syracuse, N.Y.: Division of Special Education and Rehabilitation, Syracuse University, 1967; pp.27-38.

Robertson, A. G.: Applying systems analysis techniques to the evaluation of vocational programs. Journal of Industrial Teacher Education, 1969, 6, 30-36.

Ryan, T. Antoinette: Systems techniques for programs of counseling and counselor education. Educational Technology, 1969, 9(6), 54-61.

Sacramento, Calif., County Department of Education, Regional Educational Data Processing Center. Sacrament, Calif.: Author, 1966. (available ERIC as document #ES-000-969).

Saylor, G.: National assessment: pro and con. Teachers College Record, 1970, 71, 588-597.

Shalock, H. D.: Research and evaluation: data generating activities that vary in purpose, design and output. Paper presented at the Annual Meeting of the American Educational Research Association, Minneapolis, Minn., March 1970.

Scriven, M.: The methodology of evaluation. In: Tayler, R. W., Gagne, R. M., and Scriven, M. (Eds.). Perspectives of Curriculum Evaluation. Chicago: Rand McNally, 1967; pp.39-83 (American Educational Research Association Monograph Series on Curriculum Evaluation, No. 1).

Scriven, M.: An introduction to meta-evaluation. Educational Product Report, 1969, 2, 36-38.

Simon, George B. Comments on "Implications of Criterion-Referenced Measurement." Journal of Educational Measurement, 1969, 6 (4), 259-260.

Skager, Rodney W. Objective Based Evaluation: Macro-Evaluation. Evaluation Comment (Center for the Study of Evaluation, UCLA), 1970, 2 (2), 7-10.

Smith, Ann Z., and Brecknell, Ursula C.: Accountability of ESEA projects. Education Recaps, 1969, 9(3), 1.

Sorenson, G.: A new role in education: the evaluator. Evaluation Comment, 1978, 1(1), 1-4 (UCLA Center for the Study of Evaluation of Instructional Programs).

Stake, R. E.: The countenance of educational evaluation. Teachers College Record, 1967, 68, 523-540.

Stake, R. E.: Generalizability of program evaluation: the need for limits. Educational Product Report, 1969, 2, 39-40.

St. Louis Park, Minn., Suburban School Services Joint Board. Total Information for Educational Systems, St. Louis Park, Minn.: Author, 1967. (available through ERIC as document #ES-001-447).

Stufflebeam, D. L.: The use and abuse of evaluation in Title III. Theory Into Practice, 1967, 6, 126-133.

Stufflebeam, D. L.: Evaluation as enlightenment for decision making. In: Beatty, W. H. (Ed.) Improving Educational Assessment & An Inventory of Measures of Affective Behavior. Washington, D.C.: Association for Supervision and Curriculum Development, National Education Assn., 1969.

Wallace, R. C., Jr., and Shavelson, R. J.: Program Report 103: A Systems Analytic Approach to Evaluation: A Heuristic Model and Its Application. Syracuse, N. Y.: Eastern Regional Institute for Education, 1970.

Wardrop, J. L.: Generalizability of program evaluation: the danger of limits. Educational Product Report, 1969, 2, 41-42.

Welch, W. W.: Curriculum evaluation. Review of Educational Research, 1969, 39, 429-443.

Welty, G. A.: The Logic of Evaluation. Educational Resources Institute, 1969.

Womer, F. B.: The National Assessment of Educational Progress: concept and organization. CAPS Capsule, 1970, 3(2), 1-7.

Woodbury, C. A., Jr., Jacobson, M. D., Mosher, Edith K., MacDougall, Mary Ann, and Caplan, J. R.: Research model for state educational needs assessment. Paper presented at the Annual Meeting of the American Educational Research Association, Minneapolis, Minn., March, 1970

Worthen, B. R.: Toward a tazonomy of evaluation designs. Educational Technology, 1968, 8(15), 3-9

FOOTNOTES

1. "Formative" and "summative" are merely synonyms for "process" and "product" types of evaluation, respectively.

2. The only ways in which the models still aid me as a professional evaluator are: (a) to show to a client during face-to-face program evaluation consultation what the general steps in the process are, and (b) to use as a discussion device during workshops on program evaluation.

3. Of course, these assumptions are always open to question. The goal is to choose meaningful classification schemes for the children so that these assumptions are at lease approximated. For those highly dubious about these assumptions, he should ask himself what the alternative would be to evaluating program without falling back to the case study method in and of itself.

APPENDIX A

SUGGESTED PRIORITIES

OF DISSEMINATION

OF THIS FIRST DRAFT

I. Bureau of Special Education (Dr. Ohrtman, Dr. Cogen, and Staff) and Bureau of Quality Assessment

II Special Education Experts from Teacher Training Institutions across State

III Panel of Measurement Experts from Across Nation

IV Major special education administrative personnel from IV's, private schools, and parochial schools.

V Selected groups of special education teachers.

APPENDIX B

GUIDELINES FOR PROFESSIONAL USAGE OF ACCOUNTABILITY DATA AT LOCAL OR STATE LEVELS WITH EITHER TOTAL PROGRAM EVALUATION OR INDIVIDUAL ACHIEVEMENT MONITORING

1. At no time is a teacher or administrator to feel his job is in jeopardy because his children appear to be doing "poorly" relative to some predefined criteria. The data gathered at local or state levels is for use only by those respective officials for determining whether certain programs and techniques (not people) will be discarded or modified.

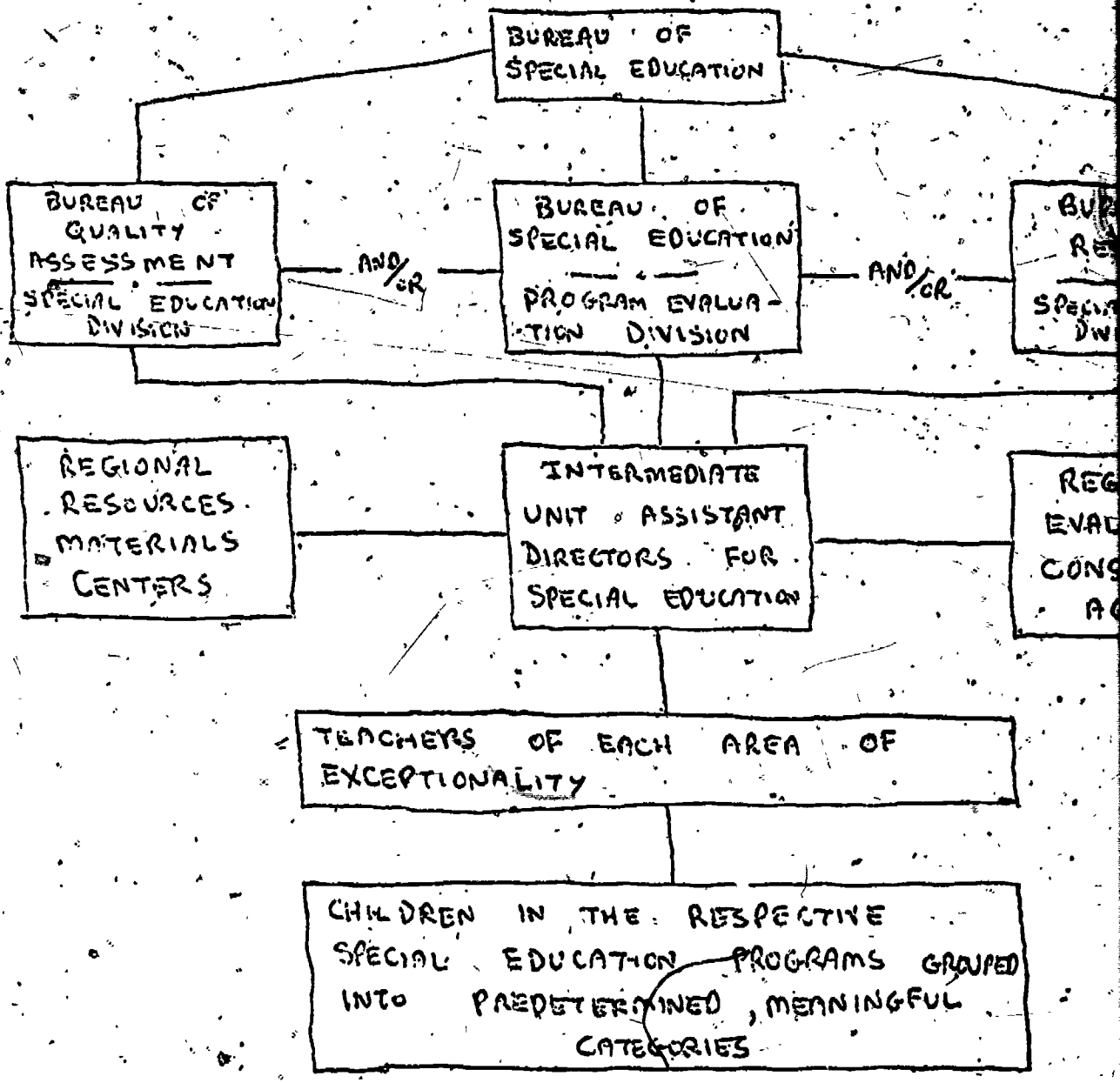
2. In any place where an accountability system is to be implemented, before a system is allowed to start, intensive in-service of all faculty (administrators and teachers) must be undertaken to avoid any misinterpretations. Complete rapport of staff with the objectives and philosophy of accountability is essential.

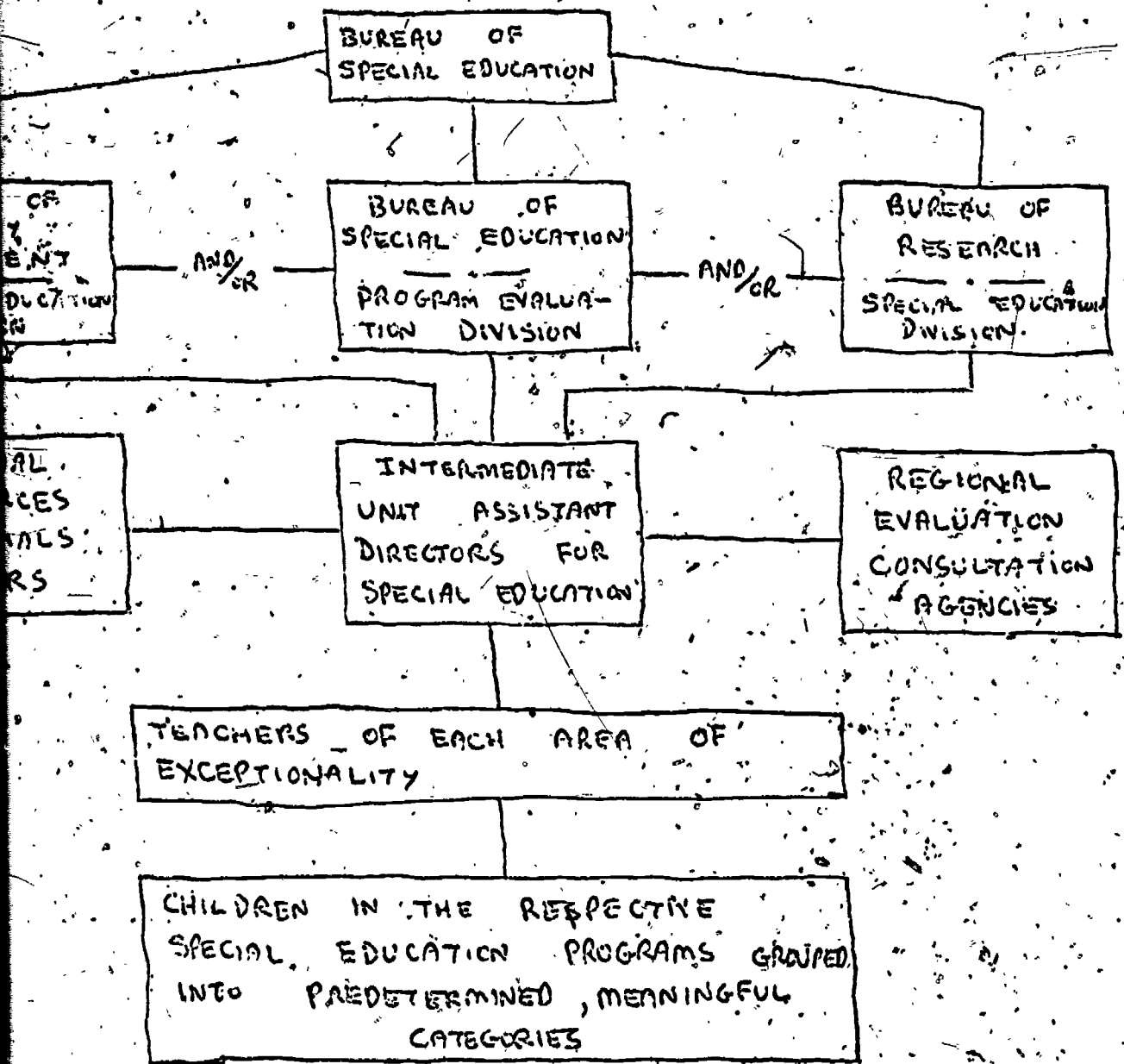
3. The state Bureau of Special Education must exert a leadership role in serving as watchdog over the use of program evaluation data at the local and state levels. The state must take appropriate action wherever misuse of data occurs.

4. Only those professionals subject to the control of the Bureau of Special Education (or those delegated by them) will have functional access to the data banks.

APPENDIX C

POSSIBLE INTERRELATIONSHIPS AMONG EXISTING AGENCIES IN CARRYING OUT A STATEWIDE FORMAL PROGRAM EVALUATION SYSTEM


BUREAU OF SPECIAL EDUCATION
BUREAU OF QUALITY ASSESSMENT SPECIAL EDUCATION DIVISION
AND/OR
BUREAU OF SPECIAL EDUCATION PROGRAM EVALUATION DIVISION
AND/OR
REGIONAL RESOURCES MATERIALS CENTERS
INTERMEDIATE UNIT ASSISTANT DIRECTORS FOR SPECIAL EDUCATION
TEACHERS OF EACH AREA OF EXCEPTIONALITY
CHILDREN IN THE RESPECTIVE SPECIAL EDUCATION PROGRAMS GROUPED INTO PREDETERMINED, MEANINGFUL CATEGORIES


BUREAU OF SPECIAL EDUCATION
BUREAU OF SPECIAL EDUCATION — PROGRAM EVALUATION DIVISION
AND/OR
BUREAU OF RESEARCH — SPECIAL EDUCATION DIVISION
AND/OR
INTERMEDIATE UNIT ASSISTANT DIRECTORS FOR SPECIAL EDUCATION
REGIONAL EVALUATION CONSULTATION AGENCIES
TEACHERS OF EACH AREA OF EXCEPTIONALITY
CHILDREN IN THE RESPECTIVE SPECIAL EDUCATION PROGRAMS GROUPED INTO PREDETERMINED, MEANINGFUL CATEGORIES

APPENDIX D

OUTLINE OF OPERATIONAL STEPS NEEDED TO IMPLEMENT A STATEWIDE FORMAL PROGRAM EVALUATION SYSTEM IN ITS FIRST YEAR

I Agree to commit any given special education program's personnel to collecting data on a regular basis -- at least twice a year; for certain types of performance, at least three times a year. (Most data will be collected by teachers.)

II Step I is accomplished, minimal common program objectives must be established that all children can be measured on. This step must be distinguished from individual pupil objectives that a highly specific educational prescription would embody. Most curriculum guides have program objectives directly or implicitly stated, altho, they are not always as operationally stated as they should be. Because program objectives are a lot easier to agree upon than are individual pupil objectives, Step II should be able to be achieved quickly and without much trouble.

III Select tests, rating scales, informal inventories, etc., that are readily available and which yield data in terms of developmental norms (developmental ages, mental ages, grade equivalents, ets.) that can be interpreted easily by workers in the field. This does NOT preclude also using locally derived measuring instruments, but for broad evaluation of program goals, commonly recognized measuring devices are best.

IV Hold in-service meetings with teachers and other special education personnel to ensure that everyone understands how to administer the instruments selected in Step III. Purposes of program evaluation system are also explained in detail to the staff. (Lack of communication between administrators and teachers is a primary source of in-process failure of many attempted programs.)

V Teachers (and, to a lesser extent, other more specialized personnel that may be required for the more "exotic" tests in the battery chosen in

Step III) administer all tests at start of year over as short a period of time as possible. Control of class atmosphere, and to a greater extent, presence of teacher aides, will be a major enabling vehicle here. It is also implied in this step that the same tests will be given at the end of the year (starting early enough before the end of the school year to allow sufficient time for everyone to be evaluated). This step, of course, with pre- and post-measures on every child, is the heart of the MINIMUM DATA-BANKING ACTIVITIES REQUIRED IN A P. E. DESIGN.

VI Thus far, Steps I through V have enabled the P.E. design to provide only raw, uninterpreted data itself. For interpretation of this data, additional machinery is required. Minimal data processing facilities should be available (as they are in a large number of I.V's). A standard format for punching data on all children on all measures onto computer cards should be arranged. Regardless of whether or not such data will be analyzed in sophisticated statistical ways, such computerized data will at least yield printouts of how each child in the program has progressed throughout the year. Even in such minimal printouts of data-banked information, a foundation for decision-making is achieved. NOTE: Duplicate computer cards of every child will be fed back to the Bureau of Special Education in Harrisburg. If all on-going programs would participate, the state would finally be able to maintain a very current picture (or "account") of what is happening throughout the Commonwealth.

VII For purposes of gaining some types of rough standards of how much progress can be expected of children with a given degree of potential (or, by the same concept, a given degree of disability in a certain area),

all children should be coded on their respective data-bank computer cards, with: (a) the degree of potential (dividing the children into three or four groups on the I.Q. continuum), and/or (b) the degree of disability on certain selected characteristics obtained on the child's latest psychological evaluation (i.e., data independent of the pre- and post-measures obtained in the P.E. design). The advantage of coding the children into certain meaningful classification is that finally program administrators will be able to pay to school boards, teachers, parents, and other groups how much progress can be expected usually with a child of given disability and/or potential. No one is yet able to provide such answers. This is one of the basic purposes of a data bank. NOTE: We are not trying to establish norms, for, say the moderately retarded. i.e., with such norms, if one carried the norm-referenced idea to completion, he could say, in all sincerety, that a child who initially tested as a moderately retarded youngster in his first psychological evaluation and who now is not demonstrating academic performance in line with expectation for such a child, is "abnormally abnormal"! The data-banking idea in special education is primarily meant to yield information with which to judge the overall success of programs.

VIII An optional step with regard to the minimum data-banking machinery described in Steps I through VII is to take the data-bank information and place it within a research design framework with the hope of comparing one programming technique with another. Up to the current step, it has been assumed that all children of a certain characterization are undergoing the same type (speaking in general program-philosophic terms) of programming approach. However, in this step,

It is recognized that some special education organizations (school systems, IV's, etc.) might wish to compare different programming techniques on the same general types of children. Such comparisons can be accomplished in this step if careful records are kept and children are assigned to different techniques in accord with good research design.