DOCUMENT RESUME

ED 253 998 EC 171 707

TITLE	PEP: Developing Criteria for the Evaluation of Protection in Evaluation Procedures Provisions. Exploring Issues in the Implementation of P.L. 94-142.
INSTITUTION	LINC Resources, Inc., Columbus, Ohio.; Research for Better Schools, Inc., Philadelphia, Pa.
SPONS AGENCY	Bureau of Education for the Handicapped (DHEW/OE), Washington, D.C.
PUB DATE	May 79
NOTE	291p.
PUB TYPE	Viewpoints (120) -- Collected Works - General (020)
EDRS PRICE	MF01/PC12 Plus Postage.
DESCRIPTORS	*Compliance (Legal); *Disabilities; Elementary Secondary Education; *Eligibility; *Evaluation Methods; Federal Legislation; Guidelines; Models; *Referral; *Student Evaluation
IDENTIFIERS	*Education for all Handicapped Children Act

ABSTRACT

Four papers focus on implementation of protection in evaluation procedures (PEP) specified in P.L. 94-142, the Education for All Handicapped Children Act. Each of the authors commissioned to develop guidelines in PEP are represented: Reginald Jones ("Protection in Evaluation Procedures: Criteria and Recommendations"); Jane Mercer ("Protection in Evaluation Procedures"); James Ysseldyke ("Implementing the 'Protection in Evaluation Procedures' Provisions of P.L. 94-142"); and Ellis Page ("Tests and Decisions for the Handicapped"). The final section presents a summary of a 2-day panel meeting (representatives of state and local education agencies, universities, and the federal government) which examined issues such as the adequacy of child evaluations for eligibility and programming decisions, and offered recommendations regarding the development of technical assistance guides for local districts concerned with PEP. A five-page reference list is included. (CL)

Exploring Issues in the Implementation of P.L. 94-142

PEP

Developing Criteria for the Evaluation of Protection in Evaluation Procedures Provisions

Department of Health, Education and Welfare
Office of Education
Bureau of Education for the Handicapped

May 1979

Published by Research for Better Schools, Inc.

This work was developed under a contract or grant with the Bureau of Education for the Handicapped, U.S. Office of Education, Department of Health, Education and Welfare. However, the content does not necessarily reflect the position or policy of BEH/USOE/HEW, and no official endorsement of these materials should be inferred.

TABLE OF CONTENTS

FOREWORD

The papers printed here were commissioned by the Bureau of Education for the Handicapped to investigate issues of quality in the implementation of the Due Process Procedural Safeguards provisions of P.L. 94-142 (Section 615 of the Education of the Handicapped Act). A panel of educational practitioners was also convened to discuss the papers and provide recommendations to the Bureau. Their comments, together with the papers, represent the most recent thinking and activities of a number of highly qualified professionals. While the views expressed in the papers are those principally of the authors, each writer has drawn upon the experiences, writings, research, and observations of various other educators in addition to their own. The care with which both the authors and the panelists shared their thoughts and ideas is obvious throughout this publication. It is our hope that this document will not only be informative, but that it will stimulate other thoughts on the evaluation of effectiveness of implementation.

Edwin W. Martin
Deputy Commissioner
Bureau of Education for the Handicapped

ACKNOWLEDGEMENTS

Many have helped BEH in conducting the Criteria Study. Special thanks are due to staff at Buffington & Associates who played an important role in arranging the panel meeting and assembling the developed papers into this monograph. Appreciation is extended to Adrienne McCollum for her overal project direction, and to Angela Edwards, Assistant Project Director, for coordination of the countless details involved in setting up the panels and producing the monograph. Acknowledgement is also made of the efforts of Frances Fuchs in assistance with the development of study questions for the panel, and P. W. Robinson for support services.

As considerable "in-house" efforts also went into this study, special thanks are also in order for State Program Studies Branch staff – Mary Kennedy, Kathleen Fenton, Lou Danielson, Pat Morrissey, and Jim Maxwell – for review of drafts of the study papers. Finally, appreciation is extended to the authors of papers, and other panel participants, for their insights and suggestions.

Linda G. Morra
Project Officer
Bureau of Education for the Handicapped

PART A

Introduction: Overview of the Study

Linda G. Morra

Bureau of Education for the Handicapped

Two of the major purposes of Public Law 94-142, the Education for All Handicapped Children Act of 1975,* are to assure that all handicapped children have available to them a free, appropriate public education and to assure that the rights of handicapped children and their parents or guardians are protected. In the educational process, evaluation of the child is necessary in order both to determine eligibility for special education and related services, and to design an individualized educational program which meets the handicapped child's unique needs. Without the evaluation process, the adequacy and appropriateness of the child's educational program would be questionable. In developing the Act, the Congress recognized the importance of child evaluation, but also delineated three major areas of concern related to the identification and classification of handicapped children (Senate Report No. 94-168, Education for All Handicapped Children Act, June 2, 1975, pp. 26-29). These concerns were (1) the misuse of appropriate identification and classification data within the educational process itself, (2) discriminatory treatment as the result of the identification of a handicapping condition, and (3) misuse of identification procedures or methods so that the child is erroneously classified as having a handicapping condition.

In response to these concerns, the Congress wrote into P.L. 94-142 (Section 615) procedural safeguards to be provided to parents in decisions regarding the identification, evaluation, and educational placement of handicapped children, as well as, specific conditions which test and evaluation procedures and materials are to meet (Section 612(5)(C)). In addition, the Commissioner of Education was directed to issue regulations which assure that state and local educational agencies establish procedures to insure that these conditions are met with regard to testing and evaluation. The Commissioner was also directed to report to Congress concerning the procedures implemented by the states to prevent erroneous classification of children. As a result of the mandates, the Bureau of Education for the Handicapped (BEH) developed and published regulations concerning protection in evaluation procedures (45 CFR Part 121a.530 - 534). The regulations are the basis of a Program Administrative Review (PAR) procedure which has been developed by BEH for monitoring implementation of P.L. 94-142, including the child testing and evaluation provisions.

THE REGULATIONS

The regulations to P.L. 94-142 provide a framework for implementation of the protection in evaluation procedures (PEP) provisions, but leave many details to state and/or local educational agency discretion. Section 121a.530 of the regulations, for example, requires that state and local educational agencies insure

*P.L. 94-142 amends Part B of the Education of the Handicapped Act, which authorizes a formula-grant program to assist states in providing free appropriate public education to handicapped children.

that testing and evaluation materials and procuderes used for the purposes of evaluation and placement of handicapped children are selected and administered so as not to be racially or culturally discriminatory. As is evidenced in the current case of *Larry P.* vs. *Riles*, however, there is considerable controversy concerning which tests meet the non-discriminatory criterion. In this case, the issue is whether California schools may use IQ tests in decisions regarding the identification, classification, and educational placement of handicapped children. Plaintiffs argue that such tests are racially and culturally biased, while the defense argues that these tests are neutral.

Section 121a.531 of the regulations requires that a full and individual evaluation of a child's educational needs be conducted prior to the initial placement of a handicapped child in a special education program. Re-evaluation of the child is to be conducted every three years or more frequently if indicated. It is Section 121a.532 which delineates specific evaluation procedures which must be conducted. These procedures include the requirements that tests and evaluation materials: (1) are provided and administered in the child's native language or other mode of communication, unless it is clearly not feasible to do so; (2) have been validated for the specific purpose for which they are used; (3) are administered by trained personnel in conformance with the instructions provided by their producer; (4) include those tailored to assess specific areas of educational need and not merely those which are designed to provide a single general intelligence quotient; and (5) are selected and administered so that the results accurately reflect the factors that test purports to measure, rather than reflecting the child's impaired sensory, manual, or speaking skills (except where those skills are the factors which the test purports to measure). In addition, no single procedure is to be used as the sole criterion for determining a child's education program, and the child is to be assessed in all areas related to the suspected disability. This section of the regulations also specifies that the evaluation is to be made by a multi-disciplinary team or group of persons, including at least one teacher or other specialist with knowledge in the area of the suspected disability. Finally, section 121a.533 of the regulations specifies procedures to be followed in interpreting evaluation data and making placement decisions.

As stated, the every day translation of the regulations into practice is the province of state and local educational agencies. At these levels, however, there are many questions to be resolved. How does one define and implement the feasibility reference concerning testing in a child's native language? What test administration procedures should be established to insure that the procedures are non-discriminatory? How should the concept of validation be defined and operationalized? The overriding question can be viewed, at the school district level, what would exemplary implementation of the protection in evaluation procedures (PEP) provisions look like? The Bureau of Education for the Handicapped is interested in assisting states by supporting the development and

dissemination of exemplary implementation procedures. State education agencies (SEAs), responsible under P.L. 94-142 for monitoring local implementation of the PEP provisions and providing technical assistance, must take the lead in developing state standards for PEP implementation. Finally, local education agencies (LEAs) must conduct their own internal evaluations of PEP implementation. The following section describes an approach undertaken to investigate the issue of quality or exemplary procedures.

THE APPROACH

It is evident that for questions concerning quality to be addressed, criteria are needed which can be used to evaluate implementation. To stimulate thought regarding definitions of quality, the BEH undertook a study in October, 1977 to explore issues of quality in implementation of four major provisions of P.L. 94-142. This monograph summarizes activities related to one of those provisions — protection in evaluation procedures. The study had two major parts. First, four papers were commissioned to provide professional judgements of quality implementation of the PEP provisions. Second, a panel of education practitioners was convened to discuss the papers and make recommendations to BEH concerning their value and use.

In conceptualizing the sutdy, it was recognized that evaluation never takes place in a vacuum; standards are always involved. Judgements of the performance of a program or procedures are measured against either explicit or implicit standards. Standards are derived from experience, knowledge, and/or values. The difficulty is that standards will vary according to whose experience, knowledge, and values serve as the basis for the standards. For example, the regulations state that tests and other evaluation materials must be selected so as not to be racially or culturally discriminatory. Criteria for the evaluation of PEP implementation would be likely to vary, however, depending on one's interpretation of this concept. Educators and psychologists have long been concerned with test bias and have developed different approaches to the issue. One such effort, for example, has been the development of culture-fair tests. Such tests attempt to represent multiple cultures, rather than any one particular dominant cultural group. While culture-fair tests have not been good predictors of school success, an advocate might establish the criterion that they be used as a supplement to traditional IQ tests in the evaluation of children from minority groups. Another approach, the fair use of tests, is based on the premise that tests are fair; it is their use which is problematic and may result in racial or cultural discrimination. An advocate of this position might establish the criterion, for example, that a minority group child be evaluated only by an examiner who is a member of the same minority group as the child.

Because a variety of standards are possible, authors were selected for this study

whose experience, knowledge, and values would tend to be disparate. Naturally, the four papers do not represent all the possible standards of quality which could be identified. They do represent, however, four different approaches to the difficult issue of quality in relation to implementation of the PEP provision.

THE PEP POSITION PAPERS

Authors were provided guidelines which first expanded on the subject of qualitative implementation of the PEP provisions. Progress in implementation was conceptualized as a continuum; conformance with the letter of the law was viewed as one end of the continuum (minimal implementation), while a full meeting of the intent or spirit of the law would form the other (maximal) end of the continuum. Authors were to use this concept of progress in implementation in developing their papers.

Secondly, the guidelines requested that authors develop criteria that would be applicable at the LEA level or to any "public agency" directly responsible for educating handicapped children. Thus, the developed criteria could be used by LEAs interested in evaluating their own progress in implementation of the PEP provisions, as well as, by SEAs in conducting their own evaluations. The guidelines further indicated that criteria which would involve the collection of data either already available or relatively accessible to LEAs at a low cost of both time and money would be most useful.

Third, authors were requested to develop criteria for determining: (1) the quality of procedures undertaken by LEAs to implement the protection in evaluation procedures provisions of the law, and (2) the effectiveness of the protection in evaluation procedures implemented by LEAs. Thus, authors of PEP position papers were to develop criteria which could be used by LEAs as approximate indicators of the extent to which PEP procedures implemented by LEAs meet both the letter and intent or spirit of the law, and the extent to which they are effective.

Fourth, authors were asked to provide a rationale or justification for their criteria. It was expected that P.L. 94-142 and its regulations would provide a base for the development of criteria. For those criteria used as indicators of maximal implementation, authors were expected to draw from theory, research findings, the Congressional Record, personal experience, or personal knowledge of current practices. Where criteria did exceed the requirements of the law and regulations, authors were to indicate that the criteria represented desirable but not mandatory standards.

Fifth, the guidelines acknowledged the interrelationship of the PEP provisions of

P.L. 94-142 with other provisions — the individualized education program provision, due process procedures, and least restricitve environment provisions. Authors were requested to restrict themselves as closely as possible to the PEP provisions.

Finally, the guidelines requested that authors of PEP position papers consider different kinds of contextual influences on LEA implementation of the provision. Variables for consideration included, for example, the urban, rural, or suburban nature of the LEA and the length of time the LEA had been implementing SEA policies similar to P.L. 94-142. Authors were to determine whether a general set of criteria for determining progress in implementation of the PEP provisions could be used in varied contexts, or alternately, whether multiple sets of criteria were needed for LEAs in different contexts.

In the initial formulation of the study, some thought was given to later development of self-study guides which could be provided as a form of technical assistance to SEAs and/or those LEAs who wanted to evaluate progress in implementation. Over time, the position papers were conceptualized as an exploratory investigation concerning the feasibility of producing self-study guides on evaluation of implementation of the PEP provisions. The papers were not to be the prototype self-study guides. From their efforts to develop criteria, however, determination of the feasibility of the task might be made.

THE PEP CRITERIA STUDY PANEL

The second part of the study involved bringing together a group largely of education practitioners to discuss the position papers and provide recommendations to BEH. More specifically, the purpose of the panel was stated as follows: To determine the feasibility of developing self-study guides which could be used by state and/or local education agencies to evaluate implementation of the protection in evaluation procedures provision of P.L. 94-142. Feasibility was defined to include topics such as field-testing and dissemination, as well as content and format of possible guides.

The panel meeting was structured into three distinct parts. First, authors presented summaries of their papers and responded to questions. Second, a large group discussion was held concerning issues related to the study. Finally, three small groups were formed to develop recommendations for BEH. For the second and third activities, study questions were distributed to panelists prior to the meeting. These questions were intended to stimulate discussion and the formulation of additional questions by panelists.

Questions for the large group session concentrated on the conceptualization of the study as presented in the guidelines for authors and also as presented by the

actual position papers. For example, a series of questions addressed the concept of progress towards implementation, and questions were posed regarding whether all of the alternative criteria generated by the authors were indicative of implementation meeting the spirit of the law. One major question asked of the group was whether, in fact, the BEH could support any further activities based on this study without giving the impression that developed standards were Federal standards. It was stressed that BEH did not want to give the appearance of sanctioning specific standards. By legislative intent; SEAs have been given flexibility in implementation.

The group then was divided into three smaller working groups to develop specific recommendations to BEH on the possible development, field-testing, and dissemination of self-study guides. Specific questions posed for these groups involved the developers of the guides, comprehensiveness of developed guides as well as field-testing and dissemination efforts, the format of self-study guides and field-testing activities, and the utiltiy of field-testing developed self-study guides. Questions were asked additionally which requested strategies for increasing utility of the guides to LEAs.

The number of panelists was intentionally designed to be small. It was felt that a small group would encourage an informal atmosphere and lively exchange of ideas. In selecting educational practitioners for the panel, emphasis was placed on representation from state and local education agencies.

The next part of this monograph presents the four position papers. As is soon evident upon reading the papers, the authors varied in their interpretations of the task and their implementing definitions of non-discriminatory assessment. The papers have *not* been reviewed to ensure that Federal statutory and regulatory requirements are accurately stated. Readers seeking to fully understand the Federal requirements are encouraged to read the regulations for Part B of the Education of the Handicapped Act (45CFR Part 121a.; published at 42FR42473; August 23, 1977; and supplemental procedures for evaluating specific learning disabilities at 42 FR65082; December 29, 1977).

PART B

Approaches to Evaluate Implementation of the Protection in Evaluation Procedures Provision of P.L. 94-142

SECTION I

Protection Evaluation Procedures: Criteria and Recommendations

Reginald L. Jones

JONES, REGINALD L. Dr. Jones is Professor and Chairman, Department of Afro-American Studies and Professor of Education (Special Education) at the University of California, Berkeley. During 1976-77 he was coordinator of the Doctoral Program in Special Education on the Berkeley Campus. He received his Ph.D. in Psychology (with a minor in special education), The Ohio State University (1959). Dr. Jones has held appointments as Professor and Director, University Testing Center, Haile Sellassie I University, Addis Ababa, Ethiopia, Professor and Chairman, Department of Education, University of California, Riverside, Professor of Psychology, The Ohio State University and faculty positions at UCLA and at Miami, Fisk and Indiana Universities, the last cited in a visiting capacity. He has written extensively in the area of P.L. 94-142 relating to the evaluation of mainstreaming programs, mainstreaming and the minority child and mainstreaming and the mildly retarded. Forthcoming is another publication on "Special Education and the Future: some questions to be answered and answers to be questioned."

OVERVIEW OF APPROACH

The purpose of the position paper is to analyze the assessment-placement process in special education and to provide a set of checks and balances which stem from this analysis.

The plan of activity is to move from the general to the specific, that is, from an overview of the entire identification-assessment-placement process to an analysis of specific elements, components, and activities in the process. Potential sources of error occur among the elements and components and it is at these levels that protections will need to be developed. The present paper, then, develops criteria for insuring protections at major points in the identification-assessment-placement process.

Fifteen models of the identification-assessment-placement process (Abeson et al., 1975; Brinegar, 1976; Carroll, et al., 1977; Harrison, 1976; National Association of State Directors of Special Education, 1976a, 1976b; Office of the Santa Clara County Superintendent of Schools, no date; Sabatino, 1976; and Tucker, 1976) were reviewed (see Appendix). A synthesis of the models/diagrams/flow charts would suggest a process somewhat as follows: First, a school related problem is identified. The problem may be one of behavior, of achievement, of appropriateness of the administrative arrangement, or some combination of the above. Second, if formal observations and or assessment are deemed necessary, permission to engage in such activities is sought from parents/parent surrogates. Third, formal observations and assessments by various specialists (e.g., school psychologists, school social workers, resource consultants, speech therapists, physicians, and others) are obtained. Fourth, planning team is constituted to integrate information received about a child and to make recommendations for further case disposition. Fifth, an instructional plan may be formulated. Sixth, follow-up is required. Obviously, not all identification-assessment-placement activities follow the above model in the order presented, but most include the components indicated, or similar ones.

In the light of the above synthesis, and of 94-142 stipulations related to evaluation procedures, it would seem that major features of the identification-assessment-placement process around which protections need to be built include (1) LEA provisions for testing/assessment, (2) communications with parents, (3) dimensions of assessment, (4) the planning and placement team, (5) adequate test use, and (6) follow-up.

Drawing from the research literature and informed opinion, a variety of criteria are developed for assessing the adequacy of activity in each of the above six areas. Each criterion item is classified into one of three groups: (1) "Required by 94-142," that is, the particular activity must be carried out as required by 94-142 regulations; (2) "Desirable," i.e. while not required, the activity, if

conducted, would be valuable in meeting the intent of 94-142 regulations; and (3) "ideal, 94-142," evaluative activities which reflect the spirit of 94-142 regulations. An example of the system in operation can be seen by reference to Table 2, which concerns parental participation in the evaluation process. Several criteria are set forth, including the following: (a) communications about the child should be written in the parent's preferred language, (b) procedures must be developed for situations in which parents deny LEA's permission to engage in evaluation activities with their child, (c) there is a designated person responsible for certifying that policies with respect to parental involvement in evaluation activities have been followed, (d) if known and/or available, parents should be provided with the names and addresses of advocates (based on type of handicap, or racial, ethnic, or other considerations), (e) mechanisms are developed for determining parental understanding of what is to be done, and probable benefits and possible negative consequences of the actions, and (f) parents are involved in determining planning and placement team membership. Using the three level scheme referred to above, evaluation criteria a and b would be classified as required to meet the letter of the law, c and d rated as desirable, and e and f as the ideal to be achieved. Obviously, judgments about actual evaluative criteria and their placement into one of the three categories is subjective. However, inasmuch as the criteria are stated in straigntforward fashion, and the evaluative categories as well, in cases of disagreements about categorization it should be possible to reach consensus on appropriate item placement. In addition to developing and categorizing the evaluative criteria, attention is also given to such matters as validity, reliability, and practicality as well as, when appropriate, the rational and empirical bases for criterion development/selection. Actual criteria for assessing components of the identification-assessment-placement process are presented in sections following.

ASSESSING THE ADEQUACY OF SCHOOL DISTRICT TESTING/ASSESSMENT PROGRAMS

Background

There is no specific 94-142 requirement that LEA's have formalized provisions for securing test/evaluation data. Virtually all LEA's, in fact have such provisions, but they are not always organized to serve 94-142 needs to maximum advantage. Moreover, as presently organized, few have built into them the kind of accountability which make them as useful for 94-142 evaluation purposes as they might be. The purpose of the present section, then, is to appraise school district testing/assessment programs in the light of 94-142 evaluation requirements. To anticipate, it is apparent that protections need to be developed not only with respect to test selection and use, but also at the structural level, e.g., the manner in which LEA's are organized to provide test/evaluation data.

TABLE 1

ASSESSING THE ADEQUACY OF SCHOOL DISTRICT TESTING/ASSESSMENT PROGRAMS

Evaluative Criterion	*Required 94-142*	*Desirable*	*Ideal 94-142*
1. An identifiable program exists in the school district for securing data on student achievement, aptitude, and other personal characteristics of students presumed to be related to instruction and/or achievement (i.e., a testing/evaluation unit).		X	
2. Provisions exist in the school district for diagnosing and prescribing with respect to the learning and adjustment needs of individual children.	X		
3. The purposes of the assessment programs are available in written form.		X	
4. The description of school district assessment programs is made available to persons in these groups:			
A. District Administrators			
1. Routinely		X	
2. Upon request			
B. Teachers			
1. Routinely		X	
2. Upon request			
C. Parents			
1. Routinely		X	
2. Upon request			
D. Interested Citizens/Community groups			
1. Routinely		X	
2. Upon request			
E. Students			
1. Routinely		X	
2. Upon request			
5. At the district or building level, there is an Advisory Committee to the standardized (group) testing program which is comprised of representatives from these groups:			
A. Administrators		X	
B. Teachers		X	
C. Parents		X	

Table 1 Continued

Evaluative Criterion	Required 94-142	Desirable	Ideal 94-142
D. Students		X	
E. The Community at large (including advocates and representatives of groups known to be interested in these matters)		X	
6. At the district or building level there is an advisory committee to the individualized testing/specialized education program(s) which is comprised of representatives from these groups:			
A. Administrators		X	
B. Teachers		X	
C. Parents		X	
D. Students		X	
E. The Community at large (including advocates and representatives of groups known to be interested in these matters)		X	
7. At the district, sub-district, or building level, the *group testing program* is organizationally aligned directly with the district's/building's instructional department/unit.		X	
8. At the district, sub-district, or building level, the *individualized testing assessment program* is organizationally aligned directly with the district's/building's instructional department/unit.		X	
9. For any single grade level data are available on student achievement in relationship to the following:			
A. The district as a whole		X	
B. School building		X	
C. Classrooms		X	
D. Major racial groups		X	
E. Sex		X	
10. There are written provisions and formal structures for utilizing group evaluation results in instructional planning at these levels:			
A. District		X	
B. Building		X	
C. Classroom		X	
11. There are written provisions and formal structures for utilizing the results in instructional planning with individual students.			X
12. Assessment personnel have been trained to work with students of diverse economic and cultural background.			X

20

Evaluative Criteria

Criteria for assessing the adequacy of LEA testing/assessment programs are presented in Table 1. Key elements of the critieria are that an identifiable program for securing test/evaluation data should be present in an LEA, that information about it should be available in written form to school personnel, students, parents, and citizens that it should be organizationally aligned with the LEA's department of instruction, and that there should be associated with it (them, if separate LEA programs exist for individualized, and group testing/assessment programs) an advisory committee comprised of representatives from teachers, administrators, parents, and students. The program would specify what tests are given to whom, when, for what purposes, and how the results are to be used.

Discussion

It might appear, at first blush, that 94-142 evaluation activities concern only individual tests and assessment procedures. Such an assumption is incorrect. In virtually all LEA's, group testing programs and individual testing programs are interrelated in the sense that results from group tests are sometimes the first level of identification of children who may need specialized services. Moreover, any composite educational picture of a given child includes group test results. The role of group testing programs in 94-142 evaluation activities, then, must be given as careful attention as is given to evaluation materials designed to be administered to individual children — a seriously overlooked fact. The problem would be less serious if careful attention were paid to the criteria for test evaluation and use presented in Table 4, but typically this is not done. Consider only a single criterion used to evaluate tests, that of reliability, which refers to the consistency of measurement. Measurement authorities note that if decisions are to be made about the achievement gains and losses of individual pupils, the reliability coefficient should be at least .90 (though obviously information useful for other purposes, e.g. group comparisons, can be obtained from tests having lower reliabilities), and parallel forms of the test should be available.

As has been noted elsewhere (Jones, 1973), it is reasonable to ask, as a purely empirical question, how many extant school achievement tests have parallel forms, and reliability coefficients of at least .90. An analysis of some 1,649 achievement tests and measures for use at the elementary school level was undertaken by the Center for the Study of Evaluation at the University of California, Los Angeles (Hoepfner, Strickland, Stangel, Jansen, and Patalino, 1970). A representative sample of their findings were those obtained for 141 standardized reading tests appropriate for use in grade 6. Of the 141 measures, 105 (74 percent) had either no parallel forms, or parallel forms with reliability coefficients of less than .70; only 7 of the 141 measures possessed reliability

coefficients of .90 or higher. Moreover, Hoepfner and his associates judged that 112 (79 percent) of the tests had norm groups which were rated as local, poor, or outdated.

While no explicit reference has been made to group tests and evaluations in 94-142 (although not stated explicitly, the language of 94-142 regulations suggests strongly that the major concern is with tests of intelligence in general, particularly those that are individually administered), it is apparent that as much attention needs to be given to the use of group tests in the assessment/placement process as to individual ones.

A critical feature of an accountability system would seem to be an advisory committee(s) to LEA testing/assessment/special education programs. Many of the issues related to 94-142 evaluations were in part stimulated by parent and community concern about special education testing, assessment, and placement practices. Advisory committees of citizens, parents, children, teachers, and administrators to LEA testing/special education programs, if made workable, would do much to demonstrate LEA willingness to respond to community demands for the reform of certain practices. Where such advisory committees have been instituted (both at the district and at the building level) they appear to have worked to the advantage both of the LEAs and the educational program (Jones, 1976).

A final issue concerns the role of testing and evaluation activities in the district's organizational structure. In the typical school district, standardized testing and assessment is conducted for several purposes: to evaluate student achievement for purposes of comparison among schools within the LEA, between LEAs or against a national standard, to meet the requirements of funded projects (e.g., Title I ESEA), for placement within regular classrooms, and for placement in special classes.

There should be an LEA philosophy about the purposes of formal testing and assessment and also an administrative structure to accommodate the philosophy. While recognizing that test/assessment results may be needed for administrative purposes (e.g. Title I evaluations, etc.), it would seem that district statements of philosophy should emphasize the view that tests and other formal assessments are administered/conducted in order to plan instructional programs for individual students and to monitor achievement and skill development — requirements which are at the heart of 94-142 evaluation procedures. The adoption of such a point of view would suggest that responsibility for 94-142 assessment activities (especially in relationship to appraisals of individual students) should reside, in the final analysis, in the LEA's Department of Instruction and not, as is sometimes the case, in LEA departments only indirectly related to instructional activities.

PARENTAL PARTICIPATION IN EVALUATION ACTIVITIES

Background

While no specific mention is made of parental participation in the evaluation process as discussed in section 121a.530-121a.534, active parental involvement in evaluation activities is implied nevertheless. For example, the requirement (121a.533) that the placement decision is to be made by a group of persons, including those knowledgeable about the child, implies that parents are to be involved. Parental participation is required in the development of IEP's however, and it is in this context that evaluation data will most likely be presented. What follows then is a discussion of procedures for ensuring parental participation in the meeting(s) in which evaluation findings are presented, most probably within the context of developing IEP's.

Carrol, Gurski, Hinsdale, and McIntyre (1977) state the benefits to parental involvement in the assessment process quite well. They observe that

> Parental involvement in the assessment process has been shown to have immediate benefit to special educators. For instance, it has become apparent that the parents of children in need of intensive assessment may be enlisted in becoming valuable sources of diagnostic information, especially with regard to the child's peer and family interactions, health and play habits, developmental history and medical history. Moreover, parents who become actively involved in the assessment process often are willing to assist in actual program implementation, thereby providing a sense of continuity between home and school. Finally, parental involvement in assessment and programming adds a new dimension to the concept of accountability in educators — the direct accountability of educators to the parents whose children they shape. In the context of culturally appropriate assessment, this accountability to parents is particularly meaningful since it implies accountability to the child's cultural and linguistic heritage as well (p. 323).

Evaluative Criteria

Evaluative criteria that might be used to assess the degree of parental participation in evaluation activities and the evaluative process are presented in Table 2. Key criteria revolve around the need, first, for LEA's to develop a rationale for parental involvement. Additional criteria relate to the need to protect confidentiality rights, to develop guidelines for securing parental permission to engage in evaluation activities, to ascertain whether in fact parents actually understand clearly that to which they give assent, including probable benefits as well as possible negative consequences, and to make provisions for securing information on parental perceptions, preferences, and expectations. Finally, provisions for ensuring accountability with respect to implementation of LEA policies concerning parental participation are made, and a detailed set of

TABLE 2

PARENTAL PARTICIPATION IN EVALUATION ACTIVITIES

Evaluative Criterion	*Required 94-142*	*Desirable*	*Ideal 94-142*
1. A rational model for involvement of parents in evaluation activities has been developed/adopted by the district.			X
2. At the district level:			
A. There are designated persons who can answer parent and community questions about evaluation activities.		X	
B. There is a designated person with authority and responsibility for certifying that district procedures for communicating with and involving parents in evaluation activities have been followed.		X	
C. Procedures have been developed for situations in which parents deny permission to engage in evaluation activities with their child.	X		
3. Prior to any written communication about child evaluation activities, parental preferences for the following are determined:			
A. The language of written communication	X		
B. The language of spoken communication	X		
C. Meeting time	X		
D. Meeting place	X		
4. In the parents preferred language and prior to the initiation of any evaluation activities, the child's parents/parent surrogates, the written notice:			
A. Is written in language understandable to the general public, i.e., be free of educational, medical, and technical jargon/language.	X		
B. Includes an explanation of due process procedures and confidentiality rights as they relate to the proposed activities.	X		
C. Contains a description of the proposed action, why it is being proposed, and the options currently or formerly under consideration.	X		

Table 2 Continued

Evaluative Criterion	Required 94-142	Desirable	Ideal 94-142
D. Provides a description of each evaluation procedure, instrument, record, or report to be used or generated during the activity.	X		
E. Informs parents of [illegible] to refuse to participate in the a[illegible]	X		
F. Provides parents wi[illegible]me(s), title(s), address(es), and telephone number(s) of person(s) who can answer questions directly related to parental concerns.		X	
G. Provides parents with the names, addresses and telephone numbers of advocacy groups having interest in problems of the same or presumed nature as that/those under consideration and who have expressed a willingness to communicate with parents on this/these matter(s).		X	
H. Informs parents of their rights to seek a third party independent evaluation, at district expense.	X		
5. Mechanisms are provided for assurances that parents demonstrate an understanding of:			
A. What is to be done			X
B. Why it is being done.			X
C. How it will be done			X
D. Probable positive benefits of procedures/actions.			X
E. Possible negative consequences of procedures/actions.			X
F. Uses to be made of the information obtained.			X
6. Written communication from parents indicating assent to request to undertake assessment activities.	X		
7. Mechanisms for demonstrating parent agreement with:			X
A. What is being done			X
B. Why it is being done			X
C. How it will be done			X
8. School districts insure that principles of confidentiality are observed by making certain that the following take place:			

Table 2 Continued

Evaluative Criterion	Required 94-142	Desirable	Ideal 94-142
A. That parents and guardians have access to "personally identifiable" record keeping systems.	X		
B. That there are defined and publicized means for the parent (and child) to find out what personally identifiable information is on record, who put it there and how it is being used.	X		
C. That means are provided to keep information that was gathered for one purpose from being used for another.	X		
D. That there is a defined and publicized procedure for correcting, amending, and challenging records.	X		
E. That school districts must assure the validity and reliability of data for their intended use and take precautions to prevent there misuse.	X		
F. That procedures for preventing possible misuse of data are available in written form.	X		
G. That a record is kept of persons who have had access to records by their name, date of access, and purpose.	X		
9. Within the context of evaluation activities, information is solicited from parents on:			
A. Their perception of the problem		X	
B. Their child's behavior in the home and community		X	
C. The child's language dominance		X	
D. Perceived strengths of the child.		X	
E. Parental expectations and goals.		X	
10. Parents are involved as full partners in all or virtually all planning and placement activities including:			
A. Determining PPT membership			X
B. Delegation of agreed upon PPT tasks			X
C. Structuring of agenda			X
D. Use of student needs as guidelines for judging programming alternatives			X
E. Suggest student subject matter needs		X	
F. Influences others to accept a specific program		X	
G. Suggest instructional methods for student.		X	

Table 2 Continued

Evaluative Criterion	*Required 94-142*	*Desirable*	*Ideal 94-142*
H. Evaluative alternative(s) from the parental perspective.		X	
I. Review the student's educational progress.			
J. Review the continued appropriateness of the student's educational program.			
K. Present information relevant to the case.		X	
L. Gather information relevant to the case.		X	
M. Interpret information relevant to the case.		X	
N. Summarize information relevant to the case.		X	
O. Encourage others to participate.		X	
P. Critique members' actions.		X	
Q. Keep group on task.		X	
R. Resolve conflicts of opinion.		X	
S. Establish meeting dates.		X	
T. Set date for review of PPT discussions.		X	
U. Finalize decision.			X
V. Set evaluation criteria for student's academic performance in the special education program		X	
11. Whether active partners in PPT activities or not parents are to be given written notification of all assessment findings and given the opportunity to react to them prior to decisions about services to be given or case disposition.	X		

Major Sources: Yoshida, R. K., Fenton, K. S., Maxwell, J. P., and Kaufman, M. J. Parental involvement in the special education pupil planning process: The school's perspective Washington: Bureau of Education for the Handicapped, Division of Innovation and Development, State Program Studies Branch. No date (c); and Carroll, A., Gurski, G., Hinsdale, K., and McIntyre, K. *Culturally appropriate assessment: A sourcebook for practitioners.* Los Angeles: California Regional Resource Center, 1977.

criteria for assessing actual degree of parental participation in planning and placement team activities given.

The first evaluative criterion states that, ideally, an LEA should develop/adopt a rational model for involvement of parents in evaluation/IEP activities. This means, simply, that LEA's should formalize their thinking on the point(s) at which parents will be involved in evaluation/IEP activities and the benefits and possible disadvantages associated with various levels of participation. Yoshida and Gottlieb (1977) have given attention to these matters and have developed a three stage model — input ("school staff gathers psychometric, academic, social, familial, and medical information required to make a decision"), process ("The case conference(s) of the placement committee considers and evluates this information") and produce ("A decision is made which provides an eligibility statement and educational plan for the student") which quite comprehensively provides a framework for determining degree and kind(s) of parental participation. Possible roles for parents in each stage are carefully examined. In the input phase, for example, 94-142 requires that parents be permission givers. Ideally, they should be involved as information and preference givers as well. Inasmuch as 94-142 requires parental involvement in IEP development they are involved in the process phase also. The benefits and costs associated with an active (or a passive) role in the process is something that a district will need to think through.

Several authors (e.g. Carroll, Gurski, Hinsdale, and McIntyre, 1977; and Yoshida and Gottlieb, 1977) suggest that the home visit may be an important source of information for use in the assessment/IEP development process. While this is true, several precautions must be taken, particularly when middle class professionals visit the homes of lower SES and racially and ethnically different clients. There are at least two potential problems: (a) securing information to evaluate, and (b) evaluating the information secured. A perspective on the latter potential problem has been given by White (1972) who writes that

> A simple journey with the white researcher *(interviewer?)* into the black home may provide us with some insight into how . . . erroneous conclusions are reached. During this visit to the black home the researcher may not find familiar aspects of white culture such as Book-of-the Month selections, records of Broadway plays, classics, magazines such as Harper's, the Atlantic Monthly, or the New York Review of Books. He might also observe a high noise level, continuously reinforced by input from blues and rhythm radio stations, TV programs, and several sets of conversations going on at once. This type of observation leads him to assume that the homes of black children are very weak in intellectual content, uninteresting, and generally confusing places to grow up. Somehow he fails to see the intellectual stimulation that might be provided by local black newspapers, informative rapping, *Jet, Ebony, Sepia*, and the Motown sound. Black children in these same homes who supposedly can't read (even preschoolers) can sing several rock and blues tunes from memory and correctly identify the songs of popular entertainers. These same

> researchers or educational psychologists listening to black speech assume that our use of nonstandard oral English is an example of bad grammar without recognizing the possibility that we have a valid, legitimate, alternate, dialect (White, 1972, pp. 43-44).

Yoshida and Gottlieb (1977) warn of the "possible danger" in the interview situation, that the parents' may describe their child's home life inaccurately. In the light of White's analysis we must be sensitive as well to the fact that the interviewer may interpret what he sees and hears inaccurately.

Some scholars believe that, to the extent possible, interviewer and interviewee racial background should be matched. In summarizing research on the interviewer-interviewee match, Weiss (1975) concluded that "current evidence suggests that, on a limited range of race related questions, matching interviewers to respondents is advisable in the cause of accuracy. But for most questions in most places at most times, a good interviewer is a good interviewer." It should be noted that Weiss generalizes from a limited range of studies, many of which are quite old (20 years or more). Changes in interracial climate during the past decade, the sensitive issues being dealt with, and the attitudes which some groups (especially racial minority group members) hold toward the possibility of special education placement (Jones and Wilkerson, 1976) suggest, particularly in dealing with special populations, that in the context of home visits, attention be given to interviewer characteristics.

It would seem, ideally, that parents ought to have an active role in the deliberations. Criteria designed to appraise parental participation are designed with this point of view in mind, i.e. active parental participation. Thus, at the point of initial contact, parents need to be informed of what is to be done, how it is to be done, and how the information is to be used. Also, they should be informed of potential positive outcomes associated with the procedures as well as possible negative consequences. Utilizing data from Yoshida, Fenton, Maxwell, and Kaufman's (no date, c) investigation of parental involvement in the special education planning process, a variety of criteria which can be used to assess parental involvement in evaluation/IEP activities are presented. It is unlikely that most LEA's will be able to respond affirmatively to all criteria listed in Table 2, which reflect ideal and maximum parental involvement and participation. However, the criteria do represent a set of guidelines which an LEA can use to organize its thinking about desired degrees of parental involvement and participation. Also the criteria can be used by the LEA to determine how well it has followed the guidelines developed.

Obviously, LEA's need models of ideal participation in evaluation/IEP activities. They also need criteria to determine the extent to which minimally acceptable objectives have been met. At the same time, LEAs must realize that a number of factors (both positive and negative) have to be considered in decisions about degree and kind of parental participation in evaluation/IEP activities. Yoshida

and Gottlieb (1977) have summarized these considerations exceedingly well. They write:

> Assuming that greater parental participation in the decision making apparatus is associated with increased fulfillment of due process guarantees, what are some of the gains and losses that can be anticipated? On the positive side, parents may not reject school placement decisions as often, thereby reducing the number of due process hearings (Kirp and Kirp, 1976). Also, parents may become more receptive and less hostile to the school's demands, especially when they are involved in placing students in special classes. Finally, parents may be taught methods for dealing with the child in the home, thus fulfilling the "home-school" team effort so often advocated. This team relationship may become necessary as parents are required to be present during the development of the individualized education plan as proposed in 94-142. However, there may be certain disadvantages that accompany parental involvement. Greater participation may also mean more opportunities for parents to observe the system, and they may conclude that schools are not operating in the best interest of their child. More importantly, the presence of parents may require major changes in the committee's handling of the case, which may affect the degree of openness with which members state opinions and suggest solutions. These costs and benefits must be weighed when defining the parent role.
>
> These efforts to increase parental involvement in determining the educational placement and programming of their child focus on legal procedures which necessarily must be followed. However, fulfilling legal criteria should not be equated with remedying the fundamental problem which due process was intended to relieve. Do these procedures result in educationally sound practices which increase the student's achievement and adjustment? Until this question is answered positively, professionals and laymen alike should be cognizant that improvement in due process procedures does not necessarily imply a concomitant improvement in educational performance among those the litigation and legislation was designed to help most — the pupils (Yoshida and Gottlieb, 1977, p. 20).

DIMENSIONS OF ASSESSMENT

Background

Adequate protection for the student requires that assessment be comprehensive. The regulations are explicit in stating that the evaluation is to be made by a multidisciplinary team or group of persons and that the child is to be "assessed in all areas related to the suspected disability, including, where appropriate, health, vision, hearing, social and emotional status, general intelligence, academic performance, communicative status, and motor abilities." In the present section evaluative criteria are set forth for determining whether in fact provisions have been made for securing comprehensive data in areas related to the child's disability.

Evaluative Criteria

The criteria presented in Table 3, drawn directly from the regulations and other sources, are meant to represent areas that might be given attention in any comprehensive assessment of a child. It is unlikely that all of the dimensions will be necessary for each child who is the subject of study, but many will be. Teacher reports on children are required by 94-142 regulations whereas direct observations of the child in the classroom are not, although the need for such observations is not ruled out. Provisions are available for securing information on the child's cognitive, language, social-emotional, educational, and physical functioning, but that related to the family and cultural environment is often not secured in part perhaps because it is difficult to obtain. A final item requires formal attention to each evaluative item in Table 3 by the planning and placement team. Such a requirement builds accountability into the evaluative process.

Discussion

As is well known, teacher reports, particularly of child behavioral and social performance may be subject to some error. The context is important if behaviors are to be interpreted accurately. The evaluative dimension related to the direct observation (by persons other than the teacher) of (a) the classroom management system, (b) student teacher interactions (c) student peer interaction, and (d) the child himself, may supplement data from the teacher and other school personnel. To be sure, there are problems with observations in the above areas. There is first the question of relevance. While classroom behavior can be categorized on a number of dimensions it should be established that the dimensions on which assessments are made are related, in some way, to meaningful aspects of the school experience. This is not always the case for observation systems, even those developed to provide information about a child's performance in the school setting. Second, once a system has been agreed upon, there is the practical problem of who will conduct the observations. It must be acknowledged that any observational system proposed will require considerably more personnel than now exist in LEA's for such purposes.

There is concern in 94-142, and justifiably so, about the use of a single instrument or procedure for acquiring data on any dimension of interest. Within the present context this concern needs to be extended to the reliability of the observations made. Obviously, if there is to be confidence in the ratings, inter-observer reliability must be obtained. This requirement again dictates additional manpower needs. Despite the fact that obtaining observation may be expensive and time consuming its value and importance should not be underestimated. By giving attention to the context in which behavior occurs and to classroom dynamics, a more adequate assessment is likely to result.

TABLE 3

DIMENSIONS OF ASSESSMENT

Evaluative Criterion	*Required 94-142*	*Desirable*	*Ideal 94-142*
1. For each 94-142 activity involving individual student appraisal, there are LEA provisions/guidelines to insure formal consideration of the following kinds of information:			
A. Teacher reports of child's	X		
1. academic performance	X		
2. behavioral performance	X		
3. social performance	X		
B. Direct classroom observations			X
1. classroom management system			X
a. evaluation of learning rate accommodation			X
b. evaluation of child's cognitive style			X
c. evaluation of curricular content			X
d. evaluation of classroom environment			X
2. Student-teacher interactions			X
a. teacher verbal reinforcement patterns			X
b. teacher non-verbal reinforcement patterns			X
c. teacher bilingual interactions			X
3. Student-peer interactions			X
a. reports of observations of child in group settings			X
b. reports of observations of child in relationship to group norms			X
4. Direct observations of child			X
a. specific target behaviors			X
b. locus of control			X
C. Educational functioning			
1. achievement in subject areas	X		
2. learning style(s)	X		
3. strengths and weaknesses			X
D. Social-emotional functioning			
1. social-psychological development			
a. attending/receiving			X
b. responding			X
c. valuing			X
d. organizing			X
e. characterizing			X
2. self-help skills			X

Table 3 Continued

Evaluative Criterion	*Required 94-142*	*Desirable*	*Ideal 94-142*
E. Physical functioning			
1. visual	X		
2. hearing	X		
3. speech	X		
4. motor/psychomotor	X		
a. gross motor	X		
b. fine motor	X		
5. medical/health	X		
F. Cognitive functioning			
1. intelligence	X		
2. adaptive behavior	X		
3. thinking processes			X
a. knowledge			X
b. comprehension			X
c. application			X
d. analysis			X
e. synthesis			X
f. evaluation			X
G. Language functioning	X		
1. receptive			X
2. expressive			X
3. nonverbal			X
4. speech			X
H. Family			
1. dominant language	X		
2. parent-child interaction			X
3. social service needs			X
I. Cultural and Social Environment	X		
1. home			X
2. interpersonal			X
3. material			X
2. There are designated LEA personnel given the responsibility for certifying that each of the above assessment dimensions (A-I) was formally considered by placement and planning teams and either utilized or rejected as unnecessary in the case under consideration. In the latter instance, a brief justification for non-solicitation/non-utilization of the evaluative dimension is given.			X

Major sources of criteria in the above table are the following: California Regional Resource Center. *Culturally appropriate assessment — a sourcebook for practitioners.* Los Angeles: California Regional Resource Center, 1977; and National Association of State Directors of Special Education. *Functions of the placement committee in special education — A resource manual.* Washington: National Association of State Directors of Special Education, 1977.

33

Fortunately there are a number of sources of information on observations, and observational systems, several of which have direct relevance to special education (e.g., Carroll, Gurski, Hinsdale and McIntyre 1977; Lambert and Hartsough, 1971; Lambert, Hartsough, and Urbanski, 1976; Urbanski, 1976; and Weinberg & Woods, 1975). Systems for Assessment of social-psychological development and cognitive functioning (with speical education relevance) have been developed as well (Bloom, 1956; Krathwohl, et al. 1964).

TEST EVALUATION AND USE

Background

Fair/non-discriminatory use of tests is at the heart of 94-142 protection in evaluation procedures. The regulations state that "testing and evaluation procedures used for the purposes of evaluation and placement of handicapped children must be selected and administered so as not to be racially or culturally discriminatory." Moreover, tests and evaluation materials must be provided and administered in the child's native language, or other mode of communication, have been validated for the purpose for which they are used, be administered by trained personnel, and be tailored to areas of specific educational need. The thread running through concern with evaluation procedures is, simply, that they be valid for the purposes for which they are used. While simple in conception, there are, it is to be regretted, a number of difficult problems of implementation. Criteria presented in Table 4 then should be useful in assessing test and assessment instruments for degree of bias.

Evaluation Criteria

Inappropriate test use is due largely to a failure to correctly apply existing standards. In this section, therefore, no new or innovative criteria for test appraisal are presented. Rather, the test user is directed to relevant aspects of Davis', *Standards for educational and psychological tests* (1974). If applied as they should be the standards will be useful in selecting tests for administration to the general population, to racial and ethnic minority groups, to the handidapped, and to preschoolers. What follows then (Table 4) are standards which can be applied to the evaluation of tests for any pupose and which can be used with any group. Nevertheless, certain principles are highlighted in Tables 5, 6, and 7 which treat, respectively, considerations related to test use with minority group/low SES populations, the handicapped, and preschoolers — populations of special interest within the context of P.L. 94-142.

TABLE 4

TEST EVALUATION AND USE: OVERVIEW

Evaluative Criterion	*Required 94-142*	*Desirable*	*Ideal 94-142*
1. SEAs and LEAs will insure that all persons using tests in connection with 94-142 activities understand that:			
A. When a test is publisned or otherwise made available for operational use it should be accompanied by a manual, which among other things, provides information required to substantiate claims that have been made for its use. (A1.) *Essential.*		X	
B. The test manual should describe fully the development of the test; the rationale, specifications followed in writing items or selecting observations, and procedures and results of item analysis or other research. (A2.) *Essential.*		X	
C. The identity and professional qualifications of item writers and editors should be described in instances where they are relevant; for example, when adequacy of coverage of a subject matter achievement test cannot appropriately or practically be measured against any external criterion. (A2.4.) *Desirable.*		X	
D. The manual should call attention to marked influences on test scores known to be asspciated with region, socioeconomic status, race, creed, color, national origin, or sex. (B1.3.) *Essential.*		X	
E. The manual should draw attention to, or warn against, any serious error of interpretation that is known to be frequent. (B1.4.) *Essential.*		X	
F. The manual should state explicitly the purposes and applications for which the test is recommended. (B2) *Essential.*		X	

[1] Letters and figures enclosed in parenthesis (A1., A2.4 etc.) represent the identification of the evaluative item as reported in Davis, F. (Editor) *Standards for educational and psychological tests*. Washington: American Psychological Association, 1974. Adjectives following the entry, e.g. Essential, Very desirable, etc. represent the importance of the item as judged by *Standards* authors.

Table 4 Continued

Evaluative Criterion	Required 94-142	Desirable	Ideal 94-142
G. The test manual should describe clearly the psychological, educational, and other reasoning underlying the test and nature of the characteristic it is intended to measure. (B3.) *Essential.*		X	
H. The test manual should identify any special qualifications required to administer the test and interpret it properly. (B4.) *Essential.*		X	
I. Where a test is recommended for a variety of purposes or types of inference, the manual should indicate the amount of training required for each use. (B4.2) *Essential.*		X	
J. The manual should draw the reader's attention to references with which he/she should become familiar before attempting to interpret the test results. (B4.3.) *Very desirable.*		X	
K. Evidence of validity and reliability along with other relevant research data should be presented in support of any claims being made. (B5.) *Essential.*		X	
L. The manual should differentiate between an interpretation applicable only to average tendencies of a group and one that is applicable to an individual within the group. (B5.4.) *Very desirable.*		X	
M. The directions for administration should be presented in the test manual with sufficient clarity and emphasis so that the test user can duplicate, and will be encouraged to duplicate, the administrative conditions under which the norms and the data on reliability and validity were obtained. (C1.) *Essential.*		X	
N. Instruction should prepare the examinee for the examination: Sample material, practice use of answer sheets or punched cards, sample questions, etc. should be provided. (C2.) *Desirable.*			X
O. Norms presented in the test manual should refer to defined and clearly described populations. These populations should be groups on whom test users will ordinarily wish to compare the persons tested. (D2.) *Essential.*		X	

Table 4 Continued

Evaluative Criterion	Required 94-142	Desirable	Ideal 94-142
P. The test manual should report the method of sampling from the population of examinees and should discuss any probable bias in the sampling procedure. (D2.1.1.) *Essential.*		X	
Q. Norms reported in any test manual should be based on well planned samplings rather than on data collected because it is readily available. Any deviation from the plan should be reported along with descriptions of actions taken or not taken with respect to them. (D2.1.2.) *Essential.*		X	
R. A test developer must provide evidence of the reliability and validity of his/her test; it is usually reported in the test manual.		X	
S. A manual or research report should present the evidence of validity for each type of inference for which use of the test is recommended. If validity for some suggested interpretation has not been investigated, the fact should be made clear. (E1.) *Essential.*		X	
T. Statements about validity should refer to the validity of particular interpretations or of particular types of decisions. It is incorrect to use the phrase, "the validity of the test;" no test is valid for all purposes or in all situations or for all groups of individuals. (E1.1.) *Essential.*		X	
U. The test user is responsible for marshalling the evidence in support of his/her claims of validity and reliability. (E2.) *Essential.*	X		
V. All measures of criteria should be described completely and accurately. The manual or research support should comment on the adequacy of a criterion. Whenever feasible, it should draw attention to significant aspects of performance that the criterion measure does not reflect and to irrelevant factors likely to affect it. (E3.) *Essential.*		X	
W. A criterion measure should itself be studied for evidence of validity and that evidence should be presented in the manual or report. (E4.) *Essential.*		X	

37

Table 4 Continued

Evaluative Criterion	Required 94-142	Desirable	Ideal 94-142
X. The manual or research report should provide information on the appropriateness of or limits to the generalizability of validity information. (E5.) *Very desirable.*		X	
Y. Validity coefficients are specific to the situations in which they are obtained. If the manual is to suggest generalization of validity for prediction of a given kind of criterion construct, it must present data suggesting the limits of generalizability regarding population or sample characteristics, situational context variables, or variations in criterion measurement. (E5.2.1.) *Very desirable.*		X	
Z. Local collection of evidence on criterion-related validity is frequently more useful than published data. (E5.2.2.) *Desirable.*			X
AA. The sample employed in a validity study and the conditions under which testing is done should be consistent with recommended test use and should be described sufficiently for the reader to judge its pertinence to his/her situation. (E6.) *Essential.*	X		
BB. Any selective factor determining the composition of the validity sample should be indicated in a manual or research report. The sample should be described in terms of those factors thought to affect validity such as age, sex, socioeconomic status, ethnic origin, residential region, level of education or other demographic or psychological characteristics. (E6.1) *Essential.*		X	
CC. Evidence of validity should be obtained for subjects who are of the same age or in the same educational or vocational situation as the persons for whom the test is recommended. Any deviation from this requirement should be described in the manual or research report. (E6.1.1.) *Essential.*	X		
DD. If a test is used for differential diagnosis, the manual should include evidence of the test's ability to place individuals in diagnostic groups rather than merely to separate diagnosed abnormal cases from the normal population. (E8.3.1.) *Essential.*	X		

Table 4 Continued

Evaluative Criterion	Required 94-142	Desirable	Ideal 94-142
EE. A test user should investigate the possibility of bias in tests or in test items. Whenever possible, there should be an investigation of possible differences in criterion related validity for ethnic, sex, or other subsamples that can be identified when the test is given. The manual or research report should give the results for each subsample separately or report that no differences were found. (E9.) *Essential.*	X		
FF. If the author proposed to interpret scores on a test measuring a theoretical variable (ability, trait, or attitude), his/her proposed interpretation should be fully stated. His/her theoretical construct should be distinguished from interpretations arising on the basis of other theories. (E13.) *Essential.*		X	
GG. A test manual or research report should present evidence of reliability, including estimates of the standard error of measurement, that permits the reader to judge whether scores are sufficiently dependable for the intended uses of the test. If the necessary evidence has not been collected, the absence of such information should be noted. (F1.) *Essential.*		X	
HH. The procedures and samples used to determine reliability coefficients or standard errors of measurement should be described sufficiently to permit a user to judge the applicability of the data reported to the individuals or groups with which he is concerned. (F2.) *Essential.*		X	
II. If two or more forms of a test are published for use with the same examinees, information on means, variances and characteristics of items in the forms should be reported in the test manual along with the coefficient of correlation among their scores. If necessary information is not provided, the test manual should warn the reader against assuming equivalence of scores. (F4.) *Essential.*		X	
JJ. Evidence of internal consistency should be reported for any unspeeded test. (F5.) *Very desirable.*		X	

39

Table 4 Continued

Evaluative Criterion	Required 94-142	Desirable	Ideal 94-142
KK. The test manual should indicate to what extent itest scores are stable, that is, how nearly constant the scores are likely to be if a parallel form of a test is administered after time has elapsed. The manual should also describe the effect of any such variation on the usefulness of the test. The time interval to be considered depends on the nature of the test and on what interpretation of the test scores is recommended. (F6.) *Essential.*		X	
2. In connection with 94-142 evaluation activities, LEA provisions and guidelines should exist to insure that test users:			
A. Have familiarity with *Standards for educational and psychological tests.* Washington: American Psychological Association, 1974.		X	
B. Possess a general knowledge of measurement principles and of the limitations of test interpretations (G1.) *Essential.*	X		
C. Know and understand the literature relevant to the test being used and the testing problems being dealt with. (G2.) *Very Desirable.*	X		
D. Have an understanding of psychological or educational measurement and validation and other test research. (G3.) *Essential.*	X		
E. Have sufficient technical knowledge to evaluate claims made in test manuals. (G3.1.1) *Very Desirable.*	X		
F. Base choice of tests or test batteries on clearly formulated goals. (H1.) *Essential.*	X		
G. Consider that different hypotheses may be different for students from different populations. (H1.2.) *Essential.*	X		
H. Are able to relate the history of research and development of the test to its intended use. (H3.) *Essential.*		X	
I. Understand that test scores used for selection or other administrative decisions about a child may not be useful for individual or program evaluation and vice versa. (H5.) *Desirable.*	X		
J. Know how to translate test results into instructional strategies.	X		

40

Table 4 Continued

Evaluative Criterion	Required 94-142	Desirable	Ideal 94-142
K. Understand potential shortcomings of tests when used with linguistically different and/or racial and ethnic minority groups.	X		
3. In connection with 94-142 evaluation activities, at the LEA level:			
A. Provisions are made to insure that those who administer and interpret tests have been trained appropriately for this responsibility.	X		
B. Procedures are established for periodic internal review of test use.		X	
C. Guidelines exist to insure that test scores are reported only to people who are qualified to interpret them.	X		
D. Programs are available to train assessment and instructional personnel to work with children of diverse racial and ethnic backgrounds.			X
E. There is a reasonable match in the district/building between the ethnic-SES student mix and instructional personnel.		X	
F. Provisions exist to insure that assessment personnel have language skills to communicate in the native language of any child subject to assessment.	X		
G. If such personnel are not available, formal provisions exist for securing such personnel or evaluation services using neighboring, state or regional resources; mere assertion that appropriate evaluation personnel are unavailable is not acceptable.			X
H. Personnel trained in tests and measurements with a responsibility to carry out validity studies at the local level, and to advise on test selection, use, and interpretation are available.			X
I. Provisions exist for securing consultant services in the above areas if no district personnel are available for such assignment.			X

41

Discussion

Criteria presented in Table 4 were drawn largely from *Standards for educational and psychological tests* (Davis, 1974). The *Standards* can be easily converted to checklists which LEAs can use to appraise any test considered for administration in the district. If LEA's are serious about fair test use then the Standards must be followed, or at least applied. To be sure, few tests, if any, are constructed well enough to meet all criteria specified by the *Standards*. Nevertheless, careful application of the *Standards* can lead to improvements in the selection and use of tests.

Special issues related to use of test instruments with minority group persons are presented in Table 5. Reasons for minority group concern about test use are widely known, and have been summarized in a number of sources (De Avila, 1976; Dent, 1976; Jones and Wilderson, 1976; MacMillan and Meyers, 1977; Samuda, 1976; and Sattler, 1975, to name a few). Details need not be belabored here, but it will suffice to note that bias is thought to exist at the content level where decisions are first made about what items to include in a test (the perspectives of minority group members are excluded), at the level of standardization, where decisions are made about the population for whom the test is appropriate, at the level of administration, in which tests are administered by persons unfamiliar with the patterns of language, behavior and customs of the person being examined, and at the level of validation where efforts are undertaken to determine whether or not the tests accomplish what they were designed to accomplish. Criteria presented in Table 5 are designed to address these issues.

There has also been concern about fair test use with the handicapped. It was noted, for example (Jones, 1973) that, for standardized achievement tests, data on reliability and validity were rarely, if ever, reported for populations of handicapped persons. Evaluative criteria presented in Table 6, therefore, point to considerations for assessing the adequacy of standardized tests proposed for use with populations of handicapped persons.

There will also be a need, under 94-142, to conduct evaluations of preschool children. In the present context, the preschool child is defined as one between the ages of 3 and 5. Three years of age is the legally mandated lower age limit for service under 94-142. Six is the age at which most children enter school. For several reasons, guidelines and protections for school age children are much better formulated than those for "preschool" children. First, there seems to be some urgency to deal with children already in school, who must be served now — least restrictive environments provided. Second, involvement in activities for the preschool handicapped is relatively new for special education and most current personnel probably have little training and background in this area. In any case, it should not be assumed that guidelines developed for use with school aged

TABLE 5

FAIR TEST USE WITH MINORITY GROUP/LOW SES POPULATIONS

Evaluative Criterion	*Required* 94-142	*Desirable*	*Ideal* 94-142
1. Examiners are specially trained to work with minority group/low SES populations:			
A. Have had coursework and/or workshops devoted to the speech, language, social, and behavorial characteristics of diverse minority group/low SES populations including that of the student being assessed.			X
B. Have had supervised experience in assessment of children from diverse minority group/low SES populations including that of the student being assessed.			X
2. Examiner expresses confidence in ability to fairly assess the child under consideration.			X
3. There are appropriately trained district assessment personnel of the same racial/ethnic/SES makeup as the child being assessed who can be consulted for assistance in and review of the assessment.			X
4. Provisions exist for external evaluations if no district personnel are adequately trained to conduct a fair assessment i.e., meet requirements of 1A, 1B, and 2 and 3 above.			X
5 For any standardized assessment instrument used, it has been determined that			
A. Minority group/low SES perspectives on item/task content were taken into account.			X
B. Minority group/low SES persons were involving in item writing or task selection.			X
C. Substantial and representative numbers of minority group/low SES persons were involved in initial item/task tryouts.			X
D. Substantial and representative members of minority group/low SES populations were included in test/instrument standardization.			X
E. Item analysis of items/tasks are available for members of different racial/SES groups.			X
F. Culturally specific items have been included, if appropriate.		X	

43

Table 5 Continued

Evaluative Criterion	*Required* 94-142	*Desirable*	*Ideal* 94-142
G. Data on the ethnic/SES applicability of norms are available.			X
5. Data are available on			
A. Validity as a function of racial group/SES membership.			X
B. Reliability as a function of racial group/SES membership.			X

TABLE 6

FAIR TEST USE WITH HANDICAPPED POPULATIONS

Evaluative Criterion	*Required* 94-142	*Desirable*	*Ideal* 94-142
1. For each test used for identifying and/or instructional planning with specific handicapped populations, SEA's or LEA's will determine that:			
A. The perspectives of diverse handicapped groups have been taken into account in test formulation.			X
B. Specific members of handicapped members are involved in item tryouts.			
C. Substantial and representative populations of specific handicapped persons are involved in test standardization.			X
D. When appropriate norms are available for specific populations of handicapped persons in the 3 - 21 age range.			X
E. Validity data are available for specific populations of handicapped persons at specific age ranges.			X
F. Reliability data are available for specific populations of handicapped persons at specific age ranges.			X
2. SEA or LEA personnel will be available to:			
A. Consult on appropriate test use with specific populations of handicapped persons.			X
B. Conduct research and development activities in the modification and/or construction of tests for use with specific populations of handicapped persons in the 3 - 21 age range.			X

45

children apply, ipso facto, to preschool ones. Many guidelines do apply and there is much commonality in guidelines for the two populations. For example, criteria for the selection of evaluation instruments, procedures for obtaining informed consent, and due process considerations apply to evaluation of preschool children as well as to school age ones. There are, however, several additional considerations that apply uniquely to evaluation of preschool populations and these need to be brought to the attention of evaluation practitioners and consumers if adequate protections are to be developed. These points are summarized in Table 7.

Finally, the regulations state that tests and other evaluation materials are to be provided and administered in the child's native language or other mode of communication, unless it is clearly not feasible to do so. The position taken in this paper is that there should be *no* conditions under which appropriate evaluation is infeasible; LEA's *must* make provisions for appropriate evaluation. Several criteria (items 3A, 3C, 3D, 3E, 3F, 3G) in Table 4, Test Evaluation and Use, are related directly to this question. The critical item is 3G which states that LEA's must make provisions for either securing appropriate personnel to conduct assessment in the home or neighboring district, or elsewhere. In the latter instance, adequate protection may require that assessments involving rare and/or difficult problems, or those involving children having unusual language backgrounds, be done at state, regional, or national centers. The development of such centers, obviously, will require appropriate efforts at LEA, SEA, and national levels.

ASSESSING PLANNING AND PLACEMENT TEAM ADEQUACY AND FUNCTIONING

Background

The Regulations require that 94-142 related evaluation be "made by a multidisciplinary team or group of persons, including at least one teacher or other specialist with knowledge of the area of suspected disability." Placement procedures regulations require that the decision is to be made by "a group of persons, including persons knowledgeable about the child, the meaning of the evaluation data, and the placement options."

Explicit in both evaluation and placement procedures, then, is the requirement that a team of individuals will be involved in deliberations about the child and his/her educational placement. While there will be variation in team composition as a function of the issues at hand, the team is expected to be multidisciplinary.

TABLE 7

EARLY DEVELOPMENTAL ASSESSMENT

Evaluative Criterion	*Required* 94-142	*Desirable*	*Ideal* 94-142
1. Early screening is limited to:			
A. These measures of organic functioning and basic, adaptive coping skills which enjoy a high degree of consensus within the health professions and affected communities.		X	
B. Those behavioral factors especially associated with learning language and speech development, motor skills and perceptual abilities.		X	
2. Specific assessment of emotional and behavioral adjustment and parent/child interaction are left to parental initiative.			X
3. The early developmental review			
A. Does not attach a label or categorize a child prior to extensive study and analysis.			X
B. Makes a dedicated effort to engage the primary caregiver, the parent, as a collaborator in the review process, and attempts to insure that the interpretation of the findings of the developmental review are culturally relevant, as well as psychologically sound.	X		
C. Recognize that there is not, at the present time, a single, universally acceptable tool for developmental review, while at the same time pointing out that there are a multiplicity of such instruments that may have practical utility in differing situations oriented toward review of individual and specified developmental functions.			X

Source: American Association of Psychiatric Services for Children, Inc. *Developmental review in the early periodic screening, diagnosis and treatment program.* Washington U. S. Department of Health, Education, and Welfare, Health Care Financing Administration, the Medicaid Bureau, April 1977.

to include regular and/or special education teachers, and to include specialists knowledgeable about the student's actual or perceived problem(s). Parents, parent surrogates, or advocates must be included as well. By whatever name (e.g. planning team, planning and placement teams, assessment team, placement committee, evaluation and placement committee, educational assessment service, school appraisal team, etc.) a multidisciplinary team is central to what is to be done, how it is to be done, and how the information gathered is to be used.

Evaluative Criteria

Criteria for appraising the effectiveness of PPT adequacy and functioning are presented in Table 8. The first evaluative item concerns the development of a framework for planning and placement activities. Such a framework would show the relationship of team activities to the LEA's instructional program, and would be organized internally to effectively discharge its mission.

Ideally, a philosophy of PPT activities and procedures should be available in written form, the thrust of evaluative criteria 1-8. Other guidelines relate to composition of PPT committees, specific PPT activities, and PPT accountability. The latter guidelines refer to such activities as insuring that a responsible LEA person be given formal authority and responsibility for monitoring PPT activities, that vehicles are developed to monitor PPT recommendations, that a written agenda be developed for each PPT meeting, and that there be a written report of the meeting's activities, analyses, conclusions, and recommendations.

Information relevant to establishing accountability in the assessment-placement process is presented in Table 9.

Discussion

The building of protections to insure that planning teams work effectively can proceed from actual knowledge of how teams operate in practice, and how they might operate, ideally, to discharge their missions. The most impressive and coherent set of analyses and findings related to planning team activities is to be found in the work of Fenton et al. (no date), Fenton, Yoshida, Maxwell and Kaufman (no date (a)), Fenton, Yoshida, Maxwell, and Kaufman, (no date (b)), Yoshida, Fenton, and Kaufman (1977), Yoshida and Gottlieb (1977), Yoshida, Fenton, Maxwell, and Kaugman (no date (a)), Yoshida, Fenton, Maxwell, and Kaufman, (no date (b)), and Yoshida, Fenton, Maxwell, and Kaufman (no date (c)). The results of research by these authors can form a background for building protections to insure adequate functioning of placement and planning teams. (It might seem, since the research was limited to only a single state, that the results should be treated with a degree of caution. However, since the processes and

TABLE 8

ASSESSING PLANNING AND PLACEMENT TEAM(PPT) ADEQUACY AND FUNCTIONING

Evaluative Criterion	*Required 94-142*	*Desirable*	*Ideal 94-142*
1. Planning and placement team (PPT) activities are developed from a rational framework.			X
2. At the district level, guidelines exist for the constitution of PPT's.		X	
3. At the district level, written guidelines exist for the conduct of PPT activities.		X	
4. At the district level, guidelines exist for contacting PPT participants.		X	
5. PPT membership include the following:			
A. A representative of the public agency, qualified to provide or supervise the provision of, special education.	X		
B. The child's teacher	X		
C. One or both of the child's parents or surrogates	X		
D. The child, where appropriate	X		
E. Other individuals at the discretion of the parents or agency			
1. Parent advocates		X	
2. Community advocates		X	
F. Evaluation personnel	X		
6. There are district provisions to insure that PPT members are informed about the team's legally assigned functions.		X	
7. Procedures are developed to insure that PPT members agree on team goals.			X
8. Written guidelines for PPT activities exist.		X	
9. A written agenda to accompany each PPT meeting is available.			X
10. PPT members are given access to all information that bears on the case.		X	

49

Table 8 Continued

Evaluative Criterion	Required 94-142	Desirable	Ideal 94-142
11. There is a single person in the district with the authority and responsibility for insuring that all information pertinent to a given case is made available to PPT's.			X
12. PPT's are given access directly to persons having information bearing on a given case, or reports from such persons are provided in written form. All oral presentations before PPT's are summarized for the record.		X	
13. Procedures are developed to insure equal status (non-specialized) participation among PPT members in all aspects of PPT activities.			X
14. For each case the potential contribution of specialists from each group listed will be *formally considered* and either *formally requested* or *formally rejected* as unnecessary.			
A. School administration		X	
B. Special education administrator		X	
C. Physician		X	
D. Parents		X	
E. School psychologists		X	
F. School social workers		X	
G. Student		X	
H. Referring teacher		X	
I. Receiving teacher		X	
J. Educational diagnostician		X	
K. Speech pathologist		X	
L. Physical therapist		X	
M. Occupational therapist		X	
N. Audiologist		X	
O. School nurse		X	
P. Guidance counselor		X	
Q. Curriculum specialist		X	
R. Methods and materials specialist		X	
S. Opthamologist/optometrist		X	
T. Vocational rehabilitation counselor		X	
U. Other specialists		X	
PPT's make certain that the child is assessed in all areas related to the suspected disability including where appropriate:			
A. Health	X		

Table 8 Continued

Evaluative Criterion	*Required 94-142*	*Desirable*	*Ideal 94-142*
B. Vision	X		
C. Hearing	X		
D. Social and emotional status	X		
E. General intelligence	X		
F. Academic performance	X		
G. Communicative status	X		
H. Motor abilities	X		
16. PPT's determine whether sufficient types of information about the student are available to it before making a decision affecting the student's instructional program.	X		
17. PPT's evaluate the educational significance of the data.	X		
18. PPT's determine the student's eligibility for special education.	X		
19. PPT's determine student placement	X		
20. PPT's formulate appropriate year-long educational goals and objectives for the student.	X		
21. PPT's develop specific short-term instructional objectives for the student.	X		
22. PPT's formally communicate with parents about changes in the student's educational program and invite response.	X		
23. PPT's formally communicate with the building administration about changes in the student's educational program and invite response.	X		
24. PPT's formally communicate with the teacher(s) about changes in the student's educational program and invite response.	X		
25. PPT's plan information needed for future review of the student's program and progress.	X		
26. PPT's make certain that each recommendation is accompanied by:			
A. A time-line for execution	X		

Table 8 Continued

Evaluative Criterion	*Required 94-142*	*Desirable*	*Ideal 94-142*
B. A statement of means by which adequacy of execution will be determined	X		
C. The specific person(s) responsible for execution of the recommendations.	X		
27. PPT's will review the continued appropriateness of the student's educational program.	X		
28. PPT's will review the student's educational progress.	X		
29. A written report of each PPT meeting will be made.	X		
30. Districts insure that guidelines exist for providing feedback to PPT participants and program implementors.		X	
31. There is available in the school district a single person with authority and responsibility for certifying that district guidelines with respect to PPT activities have been followed, and PPT recommendations carried out.		X	

Sources; Fenton, et al. (no date); Fenton, Yoshida, Maxwell, and Kaufman, [no date, (a)] Fenton, Yoshida, Maxwell, and Kaufman, [no date, (b)] National Association of State Directors of Special Education, 1976 (a); Yoshida, Fenton, and Kaufman, 1977; Yoshida, Fenton, Maxwell, and Kaufman [no date, a, b, and c]. See References for full citations.

TABLE 9
ACCOUNTABILITY IN THE ASSESSMENT–PLACEMENT PROCESS

Evaluative Criterion	*Required 94-142*	*Desirable*	*Ideal 94-142*
1. Within each LEA there are designated personnel who insure that 94-142 related evaluation/assessment/appraisal activities are:			
A. Relevant to educational needs			X
1. contain specific programming implications			X
2. contain suggestions for specific strategies			X
B. Pedagogically sound			X
C. Appropriate to the decisions to be made			X
D. Written in simple language			X
1. describe performances in descriptive terms			X
2. use nontechnical terms			X
2. Persons engaging in monitoring activities in 1A-D above shall not be participants in actual assessment-programming activities in the case under consideration. At the SEA level guidelines exist to audit and monitor 94-142 evaluation related activities.			X
3. Formal LEA provisions exist for securing appraisal of participant involvement in and reaction to the assessment-placement process.			X
A. Parents			X
1. Reactions to their own involvement degree of participation, and meeting dynamics are obtained			X
2. Degree of satisfaction with outcome is determined (with follow up as appropriate)			X
B. LEA participants (teachers, administrators, psychologists, etc.)			X
1. Reactions to their own involvement, degree of participation, and meeting dynamics are obtained.			X
2. Degree of satisfaction with outcome is determined (with follow-up as appropriate).			X

Source (item 1 above): National Association of State Directors of Special Education. *Functions of the placement committee in special education.* Washington: National Association of State Directors of Special Education, 1976.

procedures described as operational in Connecticut appear to be very much characteristic of the placement and planning activities of many LEA's, the results probably have wide generalizability).

Data from the research program on PPT functioning cited above have been drawn on heavily to develop criteria which, if operationalized, can be helpful in improving as well as monitoring PPT activities. For example, Yoshida, Fenton, Maxwell, and Kaufman (No date, b) found that program implementers, especially regular teachers often were not present at PPT meetings. They found also that there was not uniformity in communicating PPT information. Among groups receiving PPT information, e.g. regular teacher, special education teachers, and support personnel, no group received written information with consistency, and at best only 59% of one group of program implementers (special education teachers) received written and/or oral communications from PPT's; other program implementers received written communications even less frequently. Yoshida, Fenton, Maxwell, and Kaufman (no date (b)) note that "most information, except for that communicated to the special education teacher was communicated orally; documentation in the form of written communication was produced less often." Yoshida and his associates go on to suggest that "one possible method for reducing the informality of the communication network is to provide the program implementer with written documentations of the PT decisions and the file of information which was used to arrive at these decisions. Another method is to assign one PT member the responsibility for communicating with all program implementers, thereby reducing not only the number of different messages that will be transmitted but also the time commitments of PT members for communicating the PT decisions" (p. 10). Finally, the authors note that "Regardless of the method used, PT's must develop, procedures for verifying that the PT decisions and the student's program are transmitted without distortion in order to insure that the decisions arrived at with the consent of parents are the ones implemented" (p. 10). Research based observations such as those above were the basis of such Table 8 evaluative criteria as items 22-24, 26, and 29 – 31.

Fenton, Yoshida, Maxwell, and Kaufman (no date, (b)) found that "(a) not all PT's have an accurate idea about the scope of PT activities, and (b) that PT members recognized duties differently according to their roles, specifically more administrators and support personnel recognize the official PT duties than do regular education teachers" (p. 8). In yet another study, Yoshida, et al. (no date, a) found a strong positive relationship between staff role and participation in the PPT process, especially for regular and special teachers and school psychologists, in which school psychologists perceived themselves as high status and high participants, whereas teachers perceived themselves as low in status and participation. The results from this latter study were the basis for Table 8 evaluative items such as 13.

The research cited and the examples given are meant to be illustrative of possibilities for using research results to build criteria which permit an evaluation of the effectiveness of PPT functioning within the 94-142 context, and also to guide LEA's in structuring PPT activities.

When PPTs function inefficiently, or when all PPT members do not participate fully, errors of commission probably result. That is, to the extent that few specialists participate, and hence bring only limited perspectives to bear on any given case, the probability of erroneous classification is increased.

Many criteria for ideal PPT functioning have been presented. Meeting them will pose a great challenge. For example, there is no gainsaying that developing methods to assure equal-status participation among a group comprised both of professionals and non-professionals (i.e. parents) will be difficult. By putting forward the requirement of equal-status participation among PPT participants as a desirable PPT outcome (and other idealized criteria), it is to be hoped that research and programmatic activities will be stimulated to accomplish this as well as other desirable objectives, and that 94-142 implementation will be the better because of these efforts.

Follow-up

A program of follow-up would seem to have three components: (1) A timetable of activities which was developed by the planning team as part of the IEP; (2) as nearly as possible original PPT members, but in any case, the parents, and the child's regular and/or special teachers; (3) a set of guidelines which direct follow-up team composition and functioning; and (4) an LEA person responsible for certifying that all LEA guidelines for follow-up activities were met. Details of follow-up dimensions may be found in Tables 2-9.

SPECIAL ISSUES

Paradoxes in Personnel Preparation and 94-142 Implementation

The best protection in evaluation procedures is to adequately train personnel. Clearly, the definition of protection in evaluation procedures should be interpreted to mean evaluation by persons competent to engage in testing/assessment activities, to responsibly interpret the results, and to adequately plan instructional activities based on evaluation and other data. Guidelines and check lists relating to 94-142 evaluation activities, while useful palliatives, will not solve the basic and fundamental problems in personnel preparation that now

exist. These problems, which concern the adequacy of training for work with handicapped children in schools, must come through legislative changes at SEA levels, and probably from a national effort as well. For example, the requirement that children be educated in least restrictive environments also obligates colleges and universities, and state credentialing agencies, to adequately train personnel for such work. Coursework and/or competencies for work with exceptional children should be required of candidates seeking regular teaching credentials. In one large state having a competency based program for credentialing regular teachers, and otherwise gearing up to meet 94-142 requirements, "no competencies are yet mandated that deal with knowledge or experience with mildly handicapped learners for regular class teachers" (MacMillan, Jones, Meyers, 1976, p. 7). Course work and/or experiences, and competencies for work with ethnic minority groups similarly have not yet enjoyed widespread adoption. Moreover, many teachers do not possess knowledge of principles of tests and measurement. Such background, is essential to the proper use of tests and test results, particularly in instructional planning. Indeed, an early study (Goslin, 1967) revealed that less than 40 percent of teachers surveyed in a nationwide study had had more than minimal exposure (one course) to training in test and measurement techniques. A sizable proportion of teachers had never had even a single course in measurement techniques or attended a clinic at which testing was discussed. Moreover, "elementary and private secondary school teachers in particular report a lack of exposure to formal instruction in measurement with more than half of those who reported... indicating that they had never had any special training" (Goslin, 1967, p. 127). It is possible, of course, that principles of test use were acquired informally. However, no evidence was presented on this point, and one doubts that this was in fact the case.

The depth of teacher misunderstanding of tests and test use is great. Goslin (1967) reports, for example, that (1) teachers tend to view standardized tests as relatively accurate measures of a student's intellectual potential and achievements, a fact which may be true for some students, but surely must be questioned for others; that (2) teachers see the kinds of activities measured by standardized tests as important determinants of such academic success of children, and to a lesser extent, of their success in life after school; and (3) that teachers believe that considerable weight should be given to test scores, along with other measures such as school grades, in making decisions about allocating pupils to special classes, recommending students for college admission, and the like. Finally, (4) Goslin found that teachers who express confidence in the accuracy of standardized tests also feel that they measure the qualities necessary for future academic and nonacademic pursuits. These teachers also believe that the abilities measured are, to a significant degree innate, rather than learned. Further, they tend to feel that considerable weight should be given to test scores in making decisions about pupils. Teacher opinions of test use for instructional planning was not reported by Goslin but, obviously, systematic exploration of

such matters would be valuable as well.

The extent to which formal and up-to-date training in tests and measurements would modify views expressed by Goslin's respondents, is of course, unknown. It will suffice to note that many regular teachers appear unknowledgeable about testing and assessment matters, an understanding of which seems necessary for insuring protections in evaluation, and to implementation of other aspects of 94-142.

Questions need to be asked about the training of psychologists (and other personnel intimately involved in 94-142 evaluation activities) who often are key persons in planning and placement activities. Some states require no certification of school psychologists at all. In these states it would be difficult to insure that psychologists have adequate backgrounds for their work. In most states school psychologists are credentialed; some credentials are competency based. In California, as in other states there is no requirement that the school psychologist have teaching experience, although competencies are mandated which require familiarity with instructional programming. And even in states where teaching experience is prerequisite to school psychology certification, there is not the further requirement that the experience be with the population on whom assessments are to be made. Thus a psychologist may have had all his/her teaching experience at the high school level, yet be engaged primarily in educational programming with elementary school children.

Psychologists bring a number of skills to the assessment and placement process. They often have special expertise in behavior management, in interviewing and in matters of classroom climate. However, since a major focus of activity under 94-142 will be upon the use of tests and other evaluation procedures to plan individualized educational programs, it is important that attention be given to the professional qualifications of persons (psychologists and others) who prescribe or deliver evaluation services. State credentialling requirements will need to be reviewed carefully to determine the adequacy of provisions for personnel preparation to engage in such activities.

To summarize, it has been suggested that protections in evaluation procedures require personnel adequately trained to use assessment procedures for instructional planning, and other related purposes. Available evidence suggests that many regular teachers (and other personnel) may not be adequately prepared for 94-142 evaluation-related activities. It has been speculated that the deficiency may reside in SEA credentialing requirements which do not adequately mandate competencies to carry out 94-142 evaluation requirements. The extent to which this is in fact the case needs to be carefully investigated. If found to be true, appropriate corrective steps must be taken.

Testing and Assessment, Special Education Theory, and IEP Development

Testing and assessment results are closely tied to IEP development. First, the results of tests and measures are used in part, to indicate that an educational problem does in fact exist. Second, test results may be used to pinpoint areas presumed to require remediation. Third, tests will be used to determine whether the intervention activities have been successful. In order to be of value in this test-intervene retest process, assessment instruments must be sound. For example, a test must be reliable. Reliability is related to a test's validity in that the validity coefficient cannot exceed the square root of its reliability (Cronbach, 1970). Similarly, tests must not be racially, culturally, or linguistically discriminatory, requirements which are at the heart of 94-142 protection in evaluation procedures. Criteria for assessing the degree to which tests meet appropriate standards of acceptability for use with linguistically, culturally, and exceptionally different persons have already been presented in Table 5 and 6. While problems of adequacy exist, they are probably solvable, even using presently available psychometric technologies.

It may be possible to develop tests and measures which, while not racially or culturally discriminatory, do not predict any educationally meaningful performance, or provide information which facilitates the development of instructional activities. Thus, even when appropriate bias-free tests are developed, we may still be faced with more serious problems of (a) the absence of established relationships between the attributes measured and school performance, (b) and the absence of a theory (theories) of teaching-learning in special education. The two voids are, of course, closely interrelated.

In commenting on the first void (a above), Orasanu, McDermott, and Boykin (1977) remind us that

> . . . in order for a test to be useful in the description of what a child knows relative to what is to be learned, the test must offer well defined tasks which are essential components of what must be done in the performance of some complex skilled behavior, such as reading. That is, we cannot give a child a reading test until we can show that the items on the test are well defined in the test taker's eyes and that they relate to the skill we are trying to teach. This requirement presumes that a complete and adequate analysis of the target skill is available. Adequate task analysis describes what a person must do in order to perform successfully on the final task, e.g., read and comprehend a page of text; furthermore, it must identify subskills so that tests can be constructed which will monitor a child's progress on these components (Orasanu, McDermott, and Boykin, 1977).

In using tests for IEP development it is assumed that we possess valid information about the growth and development of academic and social abilities

of special populations, that we know something about the conditions under which such growth and development takes place, about the upper level of growth for various kinds of achievement for different populations of handicapped (or indeed non-handicapped) children, and that existing tests and measures are developed well enough to be sensitive to any changes that might in fact occur (Jones, Gottlieb, Guskin, Yoshida, 1978). Regrettably, we cannot say with any certainty how much growth can be expected to occur in students with various learning and/or behavioral profiles taught by method A or method B, nor can we be certain that very many existing measurement instruments are developed well enough to enable us to confidently measure pupil gains in achievement, a critical requirement for evaluating IEP effectiveness.

Moreover, Morrisey and Safer (1977) note that

> . . . to measure program/IEP's effectiveness in terms of pupil change indicators (e.g. achievement) it would be necessary to confirm that what was prescribed was implemented, and that the variance that was observed/measured could be accounted for in terms of implementation. This would be a particularly difficult charge since IEP related activities will have varying correspondence to elements of the prescribed educational plan and take up varying amounts of the instructional day. These problems, coupled with the inherent difficulties in pre-test/post-test methods of measuring/recording pupil performances, suggest that it may be methodologically difficult to assess IEP effectiveness in this way. Moreover, the precision and frequency of documentation that would be required to collect reliable data, make use of such methods prohibitive. Therefore it may be most desirable to consider multiple and varied methods of effectiveness — cost, resources, satisfaction and pupil measures. At any rate, determining appropriate measures of effectiveness will be an initial difficult task. . . (pp. 35-36.)

Theory is critical to the development of instructional activities, and it is to be regretted that so little theory of the teaching-learning process in special education is available. While all manner of tests have been used to predict various special education outcomes, only rarely has the selection of measures been guided by theoretical models or considerations which generate the basis for their selection, which predict various special education outcomes, or which explain how they function singly, or in interaction, to lead to some specified educational accomplishment (Jones, 1978). It is difficult to see how tests and other evaluation procedures can be used effectively in developing IEP's in the absence of such knowledge.

In the context of their discussion of competency based teacher education in special education, Semmel, Semmel, and Morrisey (1976) stated the need for theory quite well. They noted that

> . . . Theoretical conceptions must seek to identify those instructional and pupil characteristics which most probably relate to pupil growth. This implies more than the construction of hypotheses related to the effects of one type of

> ...dministrative arrangement over another. What is needed are efforts to construct models which suggest that teachers with specified characteristics, who demonstrate specified observable teacher behaviors, with pupils having specified learning characteristics will produce desired pupil outcomes within the limits of specified educational contexts. The complexity of searching for functional relationships between presage, process, and product variables in the study of teacher behavior demands a sizable effort. . . Theory is a powerful tool for organizing such an endeavor. It is, to be sure, not the only promising strategy for uncovering meaningful relationships between teacher behavior and pupil growth. But it is, in our opinion, a necessary component of a total effort. . . (Semmel et al., 1976), pp. 200-201.

It should be noted, in summary, that adequate protection in evaluation procedures should refer as much to insurance that any tests and assessment procedure be valid for the development and assessment of instruction as to the requirements that they be free of racial, ethnic and SES bias. With so much justifiable concern about the racial/ethnic bias of tests, too little attention has been given to the tests' educational validity. Obviously, if evaluation procedures are to be validly effective for use in the development of IEP's, much theoretical and research work will have to be done.

APPENDIX

Models of the Identification, Assessment, Placement Process

Abeson, A., Bolick, N. & Hass, J. Evaluation and Placement

Brinegar, L. Identification, Assessment, and Instructional Planning

Carrol, A., Gurski, G., Hinsdale, K., & McIntyre, K. CRRC CAA Summary Chart

Harrison, D. B. Student Performance Inventory

Harrison, D. B. Service Delivery System

National Association of State Directors of Special Education (1976a). Flow-chart: Placement Committee Functions

National Association of State Directors of Special Education (1976b). Sequence of Activities Program Model

National Association of State Directors of Special Education (1976b). Sequence of Identification Activities-Program Model

National Association of State Directors of Special Education (1976b). Sequence of Assessment Activities-Program Model

National Association of State Directors of Special Education (1976b). Sequence of Placement Activities-Program Model

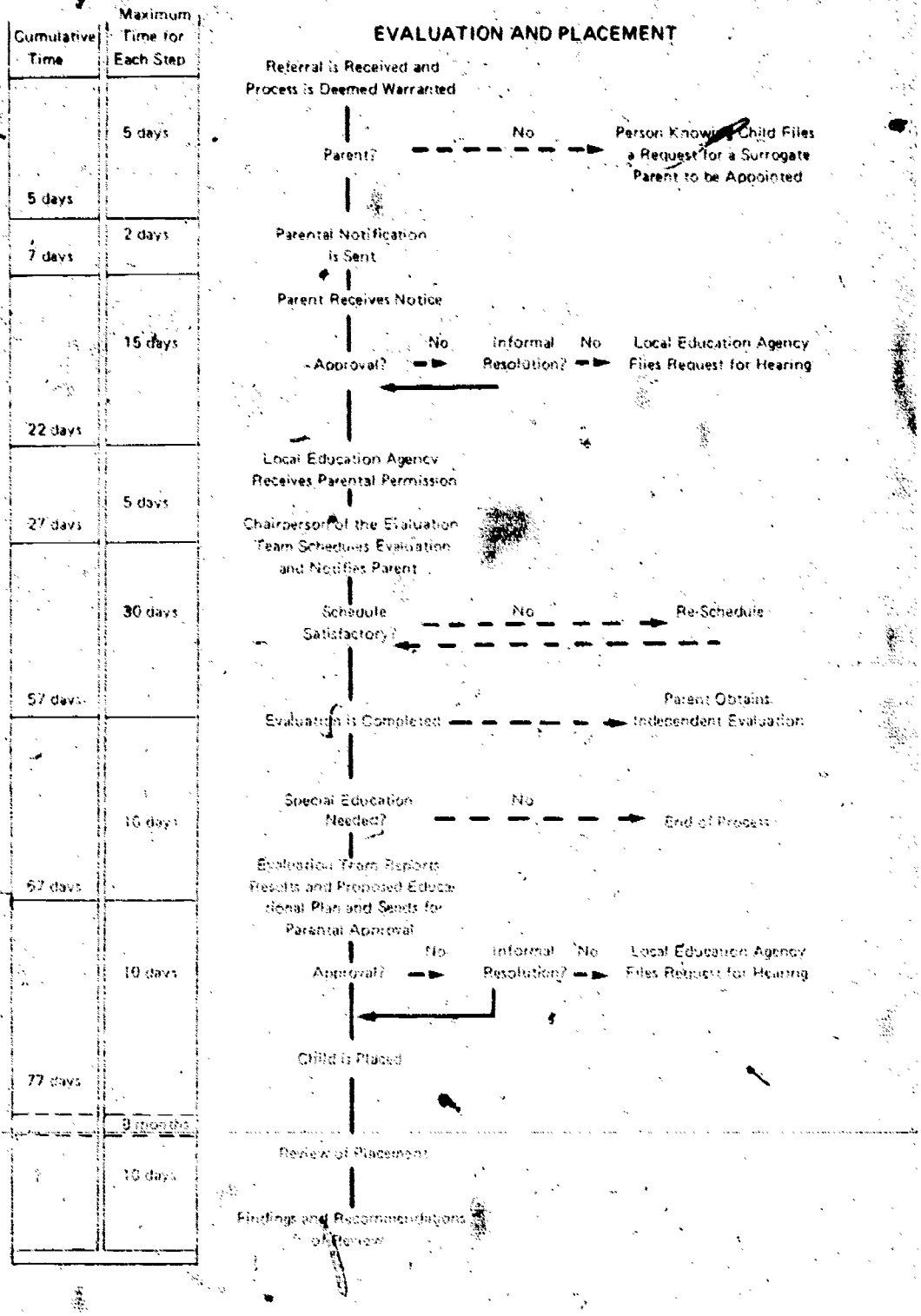
National Association of State Directors of Special Education (1976b). Sequence of Instructional Activities-Program Model

National Association of State Directors of Special Education (1976b). Sequence and Timing of Events-Program Model

Office of the Santa Clara County Superintendent of Schools. Procedural Due Process Safeguards

Sabatino, D. Six Step Sequential Model for N.R.R.C.P. Rural Unit

Tucker, J. A. Comprehensive Individual Assessment for Possible Mildly Handicapping Conditions



Source: Abeson, A., Bolick, N., and Hass, J. A. *A primer on due process.* Reston, Virginia: The Council for Exceptional Children, 1975.

Cumulative Time	Maximum Time for Each Step
10 days	10 days
40 days	30 days
45 days	5 days

REQUEST FOR A SURROGATE PARENT

Local Education Agency Receives Request

→ Sends Copies to State Education Agency and State Standing Board

Local Education Agency Investigates

Unable to Locate Parent? — No → Request is Dropped

Local Education Agency Requests State Education Agency Determine if There is Need for Surrogate

State School Officer or Designee is Assigned

Officer Weighs Evidence

Officer Decides and Informs State Education Agency, Local Education Agency, and State Board

Child Needs Surrogate? — No → End of Process

State Education Agency Assigns Surrogate to the Child

Cumulative Time	Maximum Time for Each Step
5 days	5 days
15 days	10 days
25 days	10 days
65 days	40 days
75 days	10 days

HEARING PROCESS

Local Education Agency Files Request → Local Education Agency Mails Notice to Parents

State Education Agency Receives Request

Parent Obtains Independent Evaluation

State Education Agency Appoints Hearing Officer

Hearing Officer Mails Notice of Hearing Schedule

Okay? — No → Re-Schedule

Hearing is Held

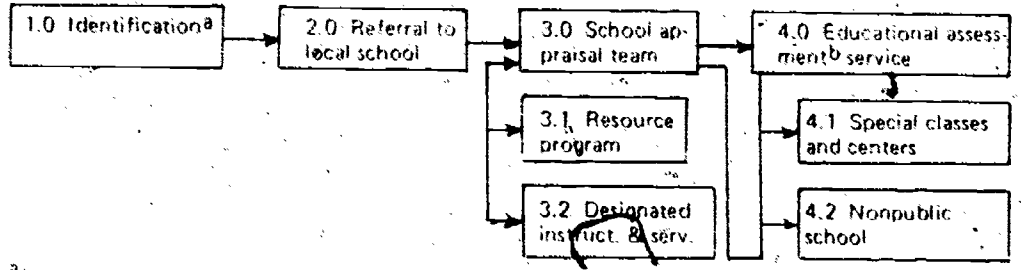
Hearing Officer Issues a Decision

Parent Approval? — No → Appeal is Initiated for Judicial Review

Issue is Resolved

63

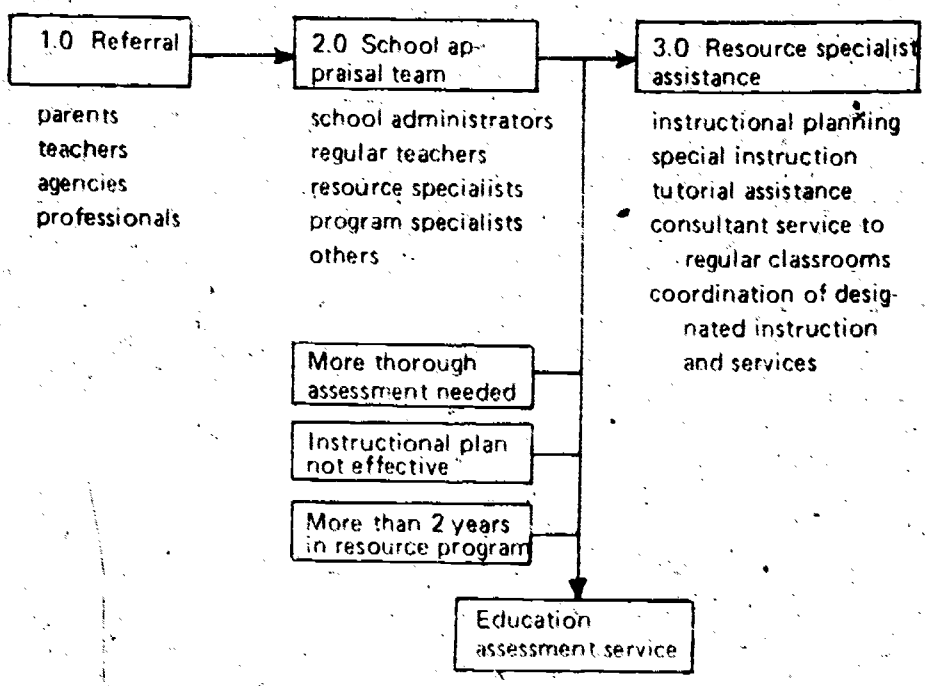
FIGURE 2
IDENTIFICATION, ASSESSMENT AND INSTRUCTIONAL PLANNING



[a]System to identify must be established (in comprehensive plan), and must include preschool, parents, teachers, and agencies.

[b]Assessment must include cognitive, affective, and sensory motor functioning.

FIGURE 3
RESOURCE SPECIALIST PROGRAM (IN EACH SCHOOL)

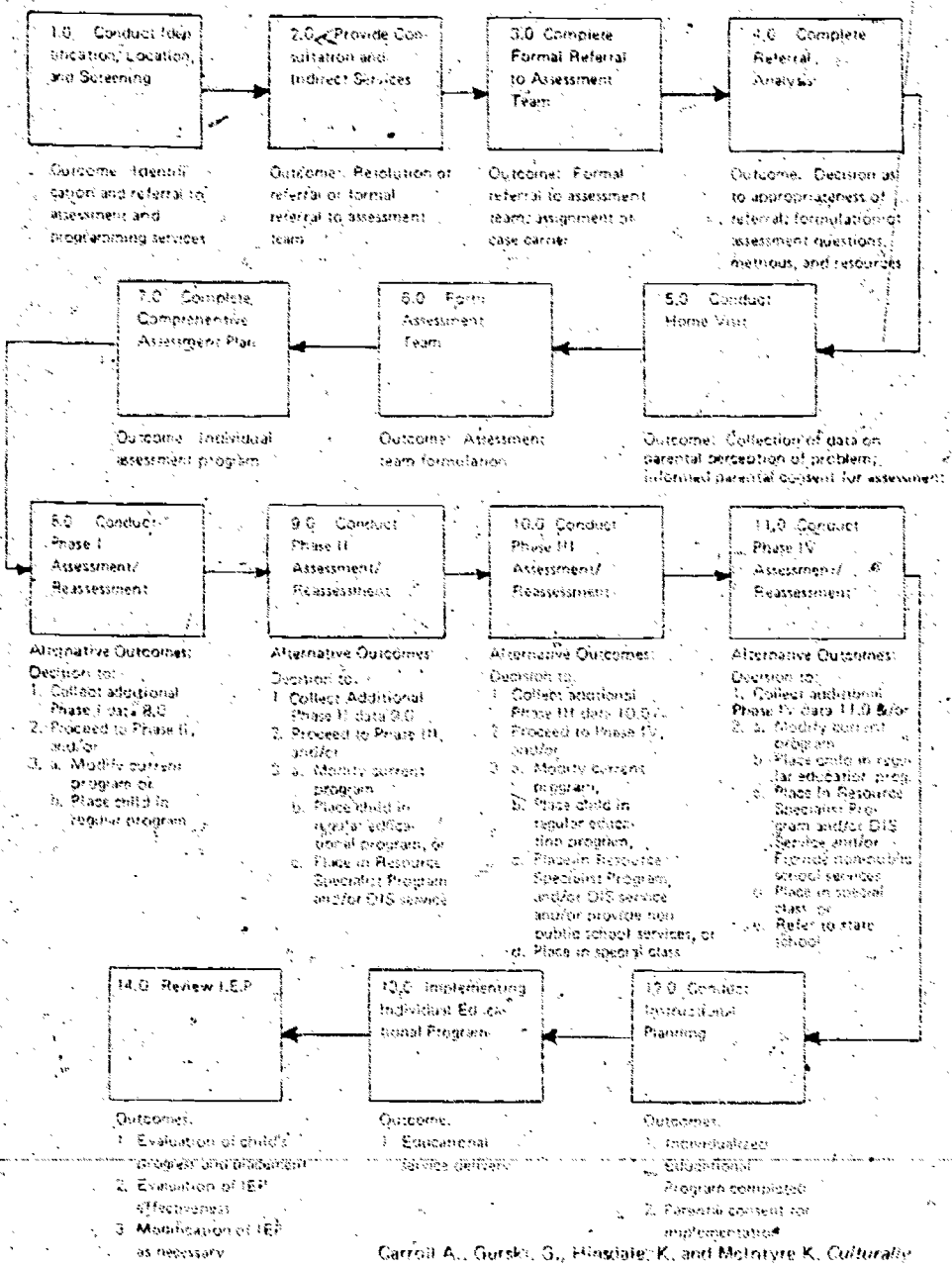


If a child needs a comprehensive assessment, there is a second level called the educational assessment service (see Figure 2) at a district or countywide level consisting of highly specialized individuals. Any child who needs this comprehensive assessment will be referred beyond the school appraisal team.

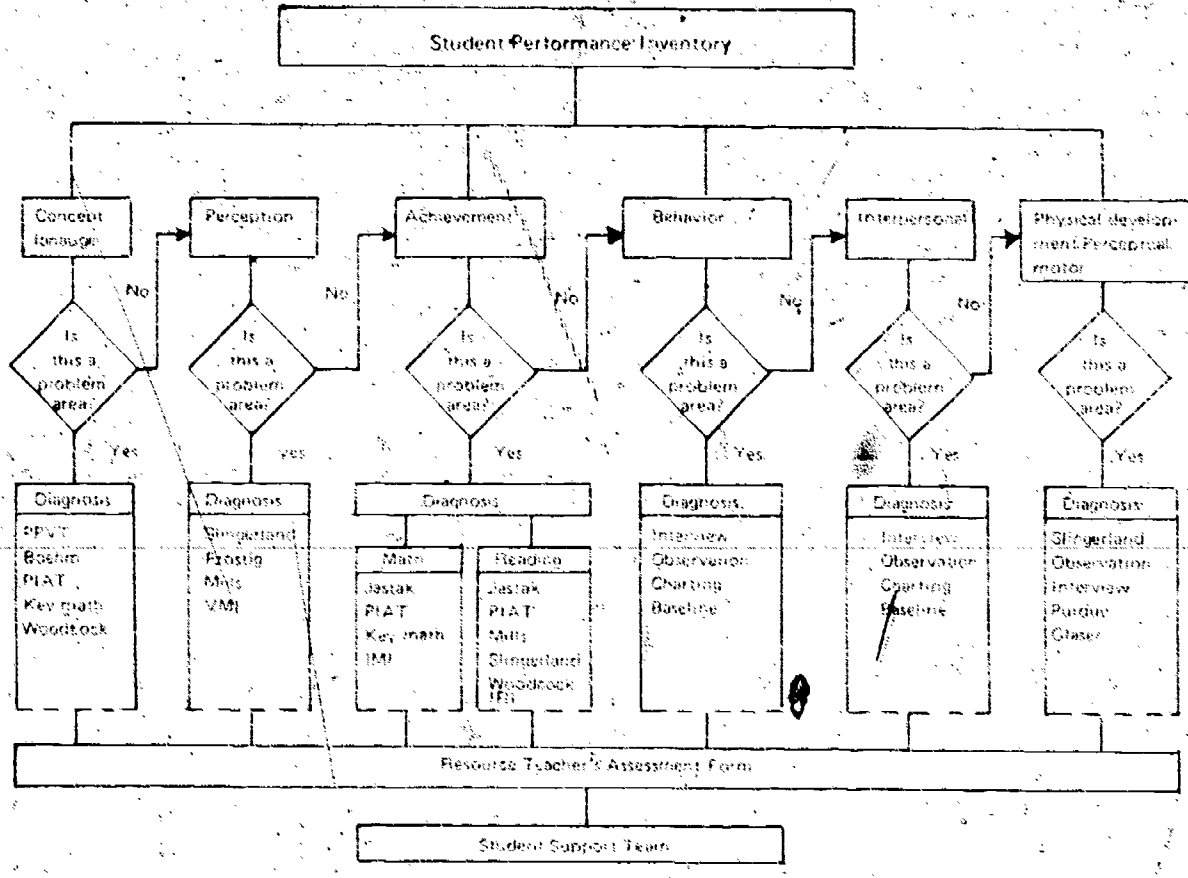
Source: Brinegar, L. Partners in learning: focus of the California master plan for special education. In J. Jordan (Editor), *Teacher, please don't close the door*. Reston, Virginia: The Council for Exceptional Children, 1976.

65

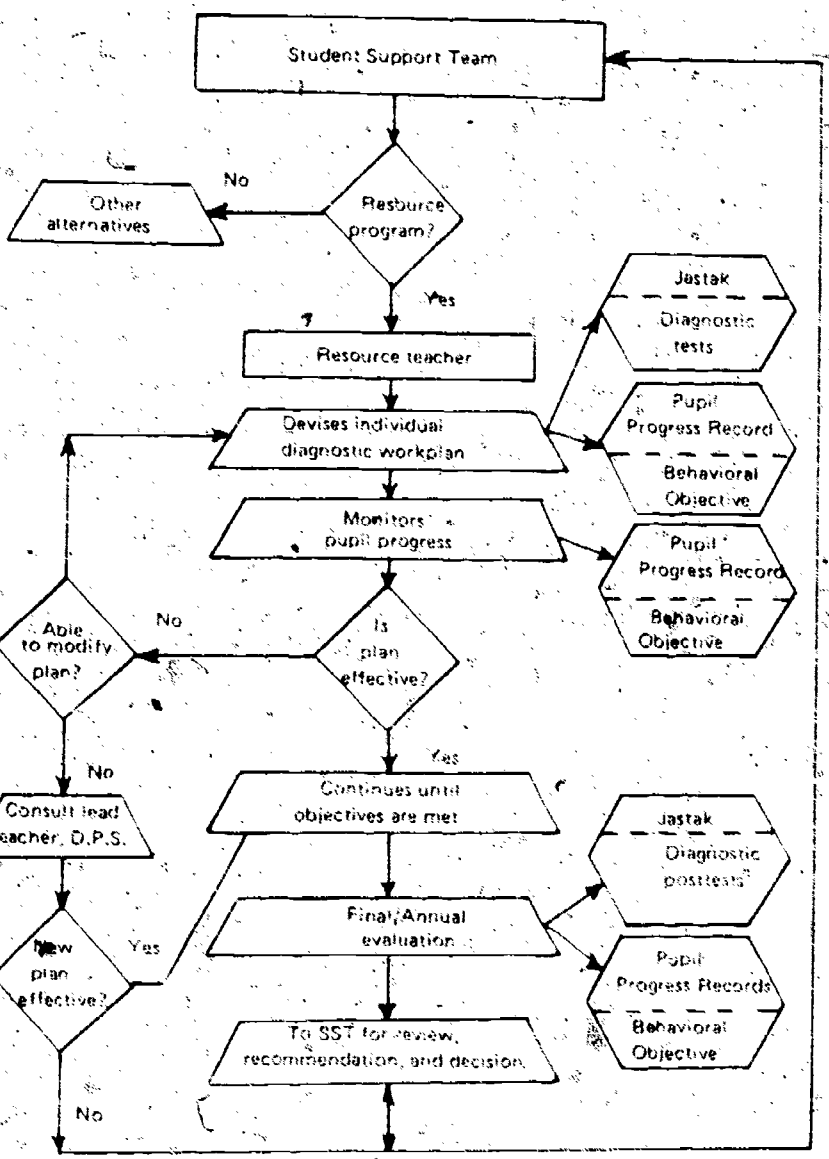
CRRC CAA PROCESS SUMMARY CHART



Carroll A., Gurski, G., Hingsale, K., and McIntyre K. *Culturally appropriate assessment: A sourcebook for practitioners.* Los Angeles: California Regional Resource Center, 1977.

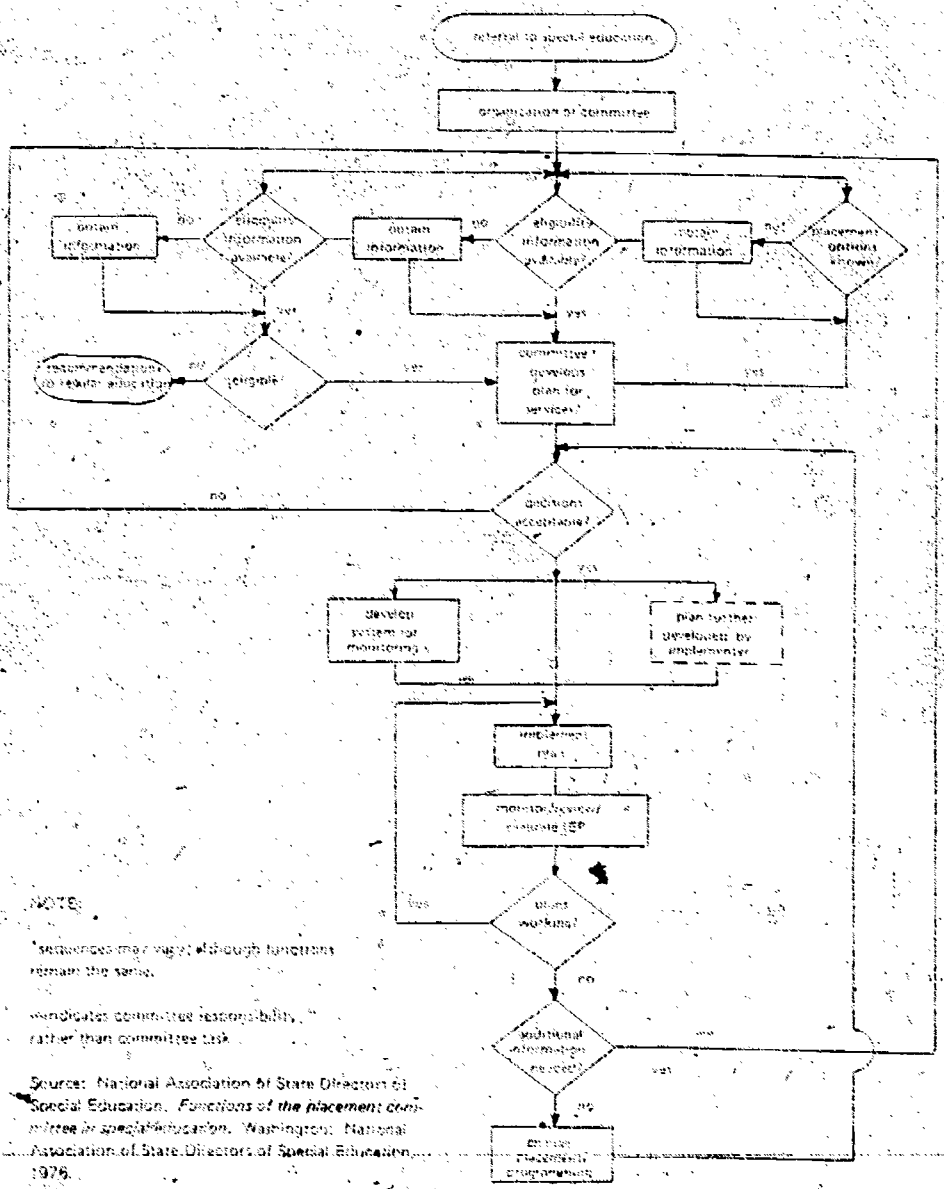


Source: Harrison, D. B. The resource teacher. In J. Jordan (Editor), *Teacher, please don't close the door*. Reston, Virginia: The Council for Exceptional Children, 1976.



Source: Harrison, D. B. The resource teacher. In J. Jordan (Editor), *Teacher, please don't close the door.* Reston, Virginia: The Council for Exceptional Children, 1976.

FLOWCHART: PLACEMENT COMMITTEE FUNCTIONS



NOTE:

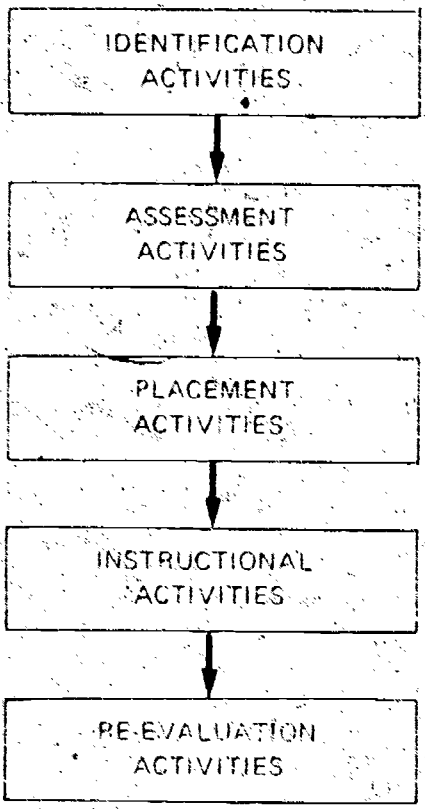
* sentences may vary; although functions remain the same.

indicates committee responsibility, rather than committee task.

Source: National Association of State Directors of Special Education. *Functions of the placement committee in special education.* Washington: National Association of State Directors of Special Education, 1976.

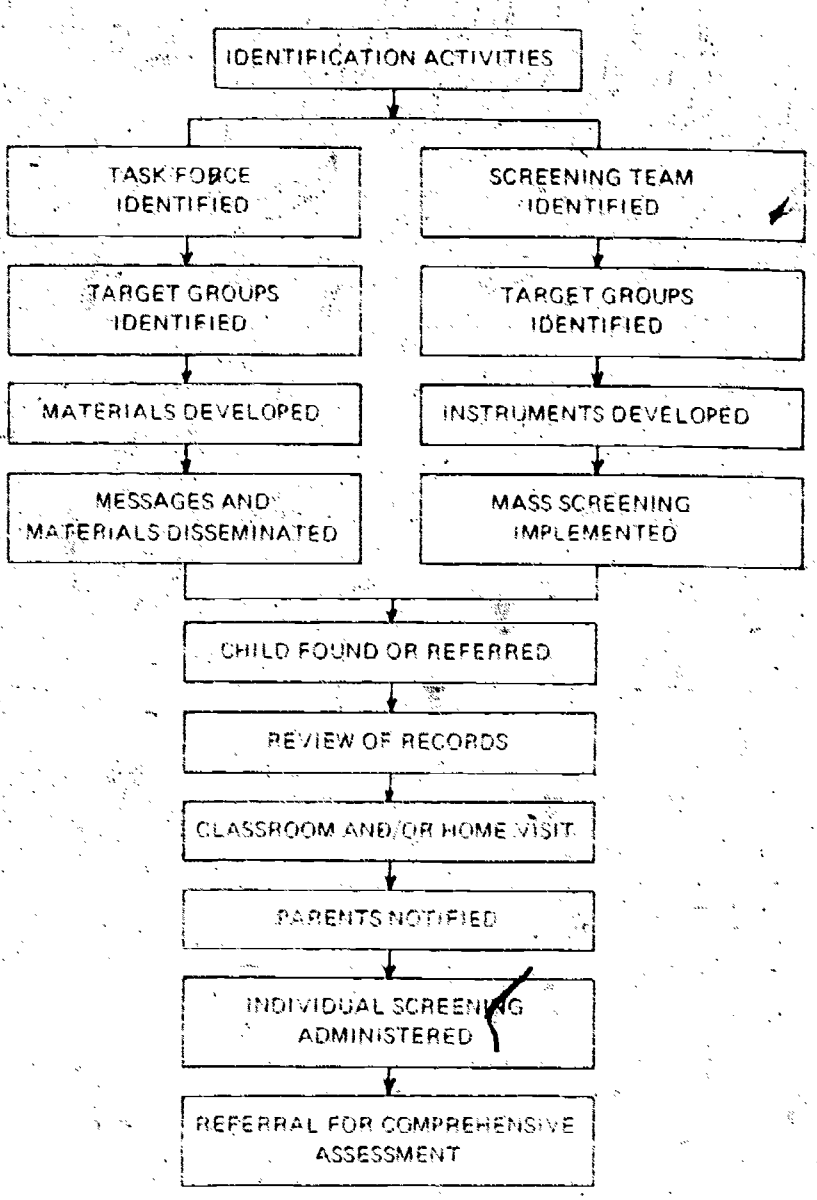
69

SEQUENCE OF ACTIVITIES
PROGRAM MODEL



Source: National Association of State Directors of Special Education. *The Prince William model. A planning guide for the development and implementation of full services for all handicapped children.* Washington: National Association of State Directors of Special Education, 1976.

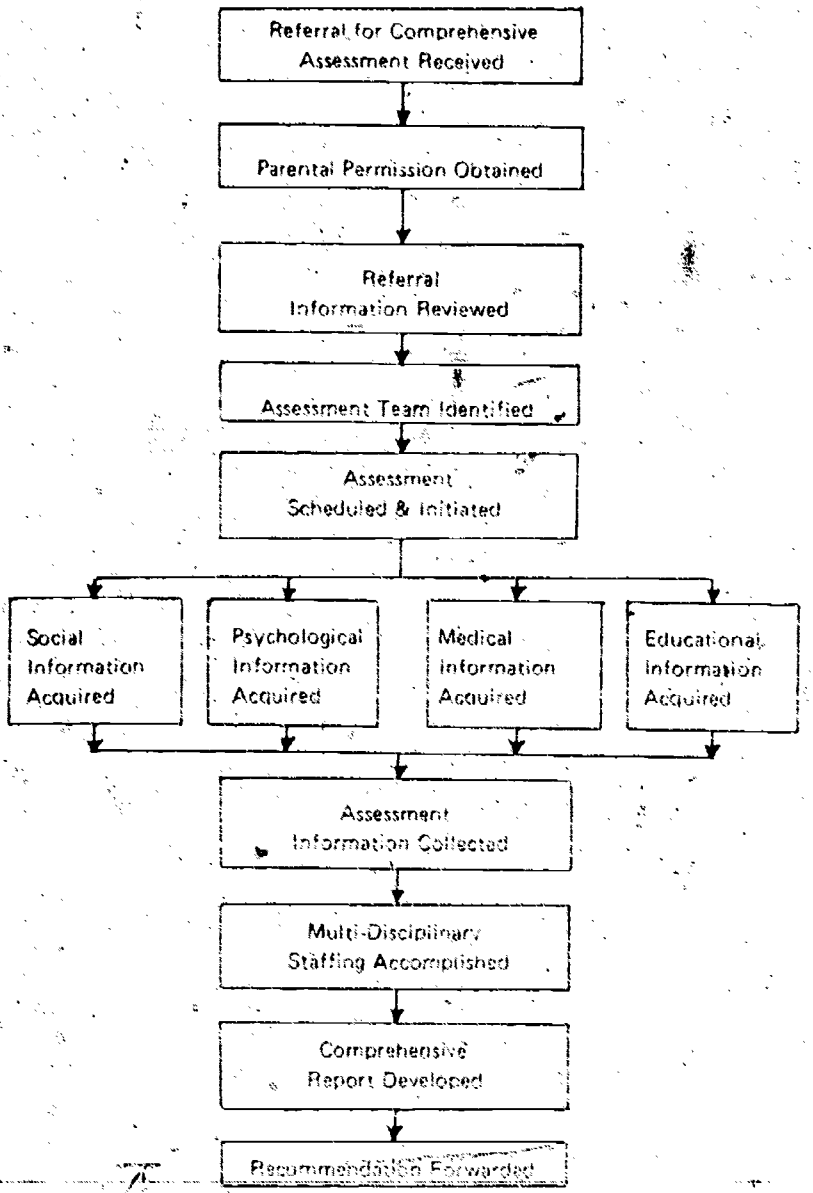
SEQUENCE OF IDENTIFICATION ACTIVITIES PROGRAM MODEL



Source: National Association of State Directors of Special Education. *The Prince William model. A planning guide for the development and implementation of full services for all handicapped children.* Washington: National Association of State Directors of Special Education, 1976.

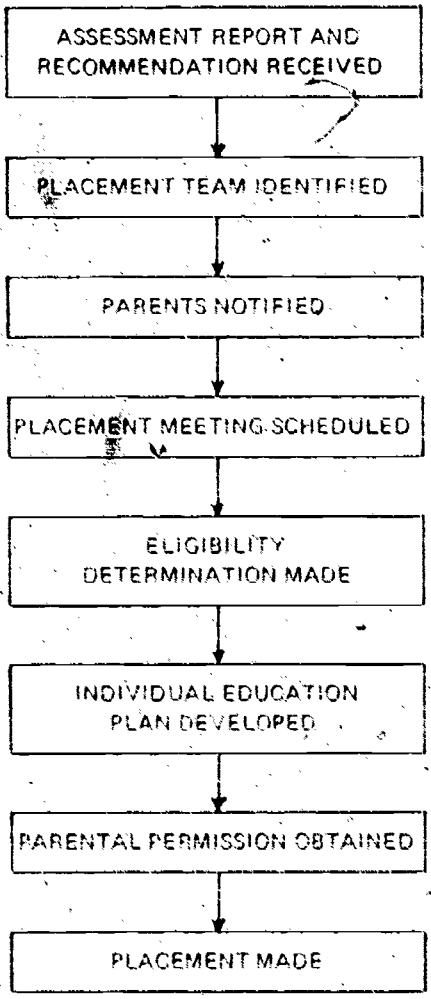
71

SEQUENCE OF ASSESSMENT ACTIVITIES
PROGRAM MODEL



Source: National Association of State Directors of Special Education. *The Prince William model. A planning guide for the development and implementation of full services for all handicapped children.* Washington: National Association of State Directors of Special Education, 1976.

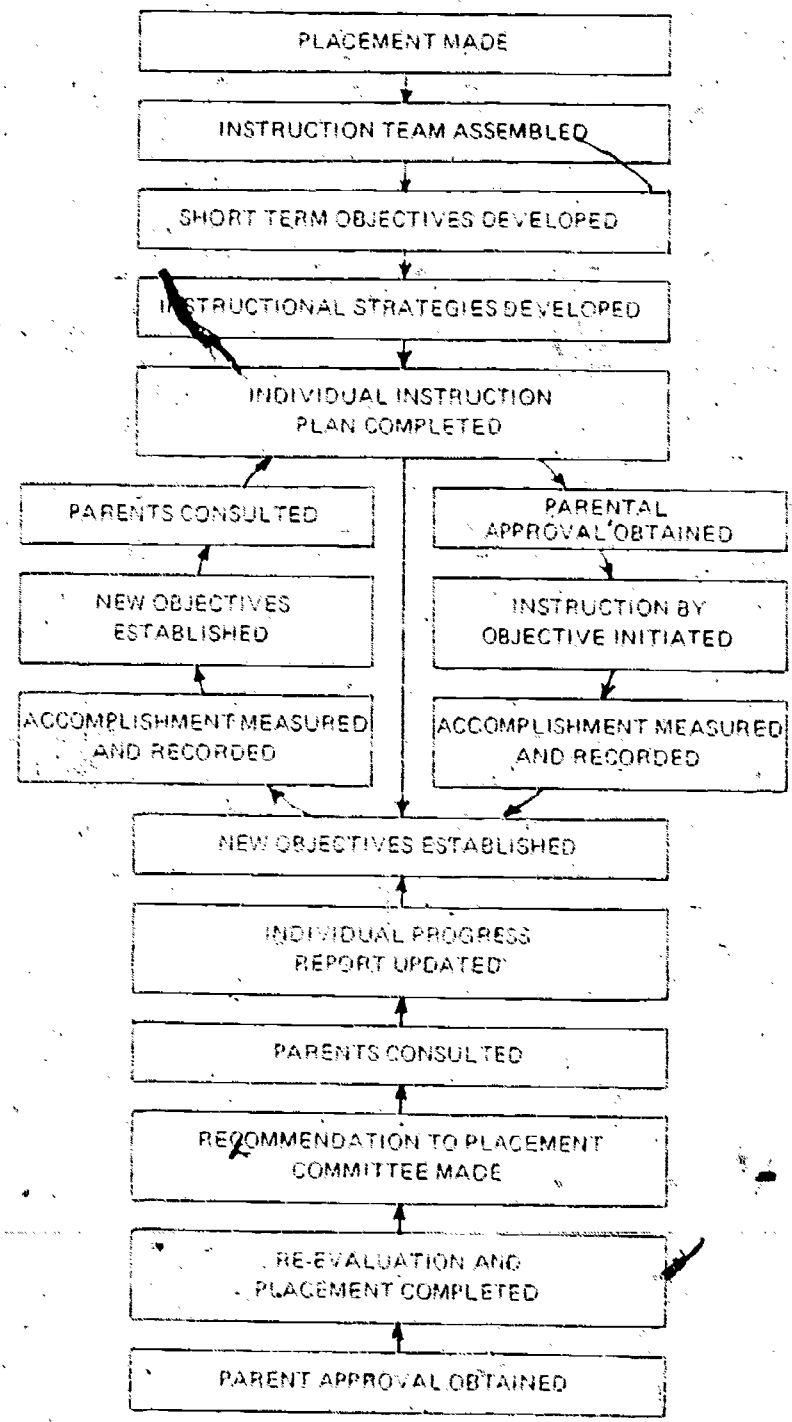
SEQUENCE OF PLACEMENT ACTIVITIES PROGRAM MODEL



Source: National Association of State Directors of Special Education, *The Prince William model. A planning guide for the development and implementation of full services for all handicapped children*. Washington: National Association of State Directors of Special Education, 1976.

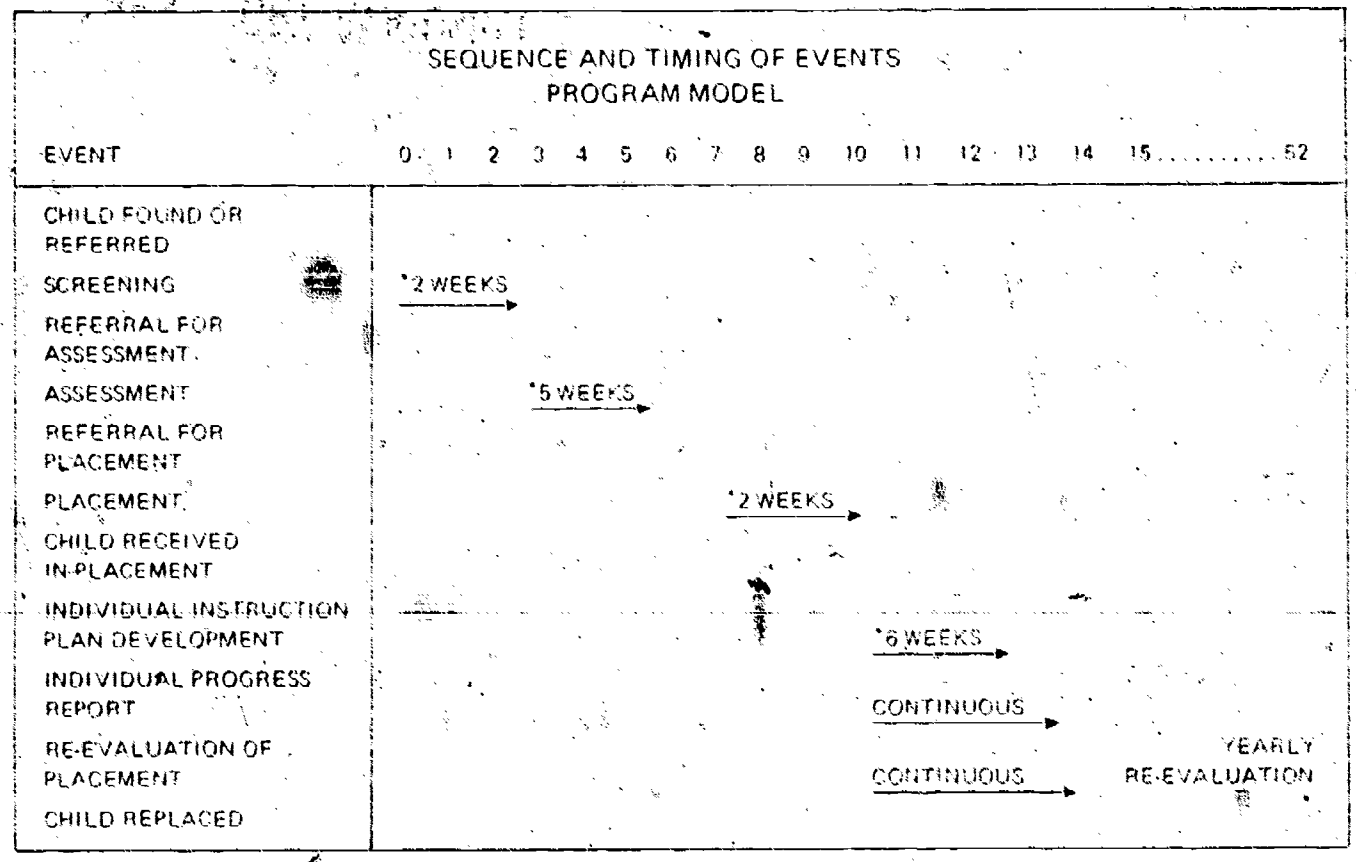
73

SEQUENCE OF INSTRUCTIONAL ACTIVITIES PROGRAM MODEL



Source: National Association of State Directors of Special Education. *The Prince William model. A planning guide for the development and implementation of full services for all handicapped children.* Washington: National Association of State Directors of Special Education, 1976.

SEQUENCE AND TIMING OF EVENTS
PROGRAM MODEL

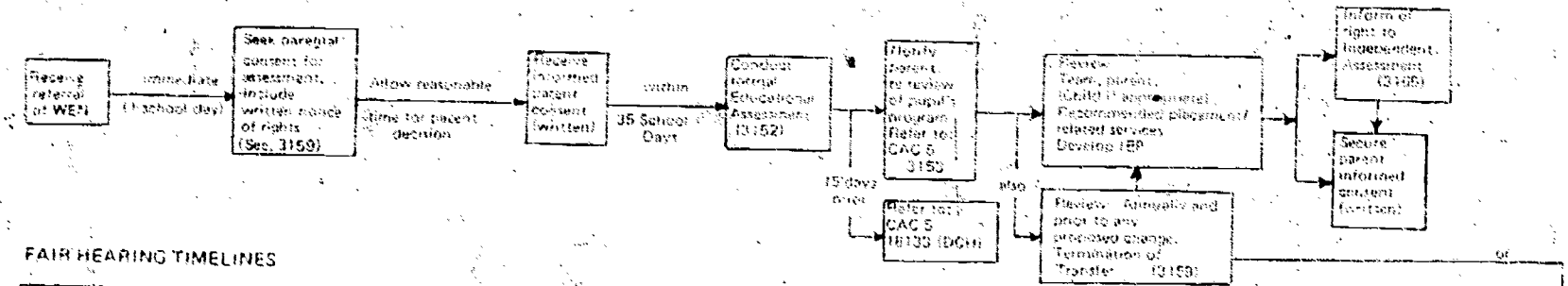


Source: National Association of State Directors of Special Education. *The Prince William model. A planning guide for the development and implementation of full services for all handicapped children.* Washington: National Association of State Directors of Special Education, 1976.

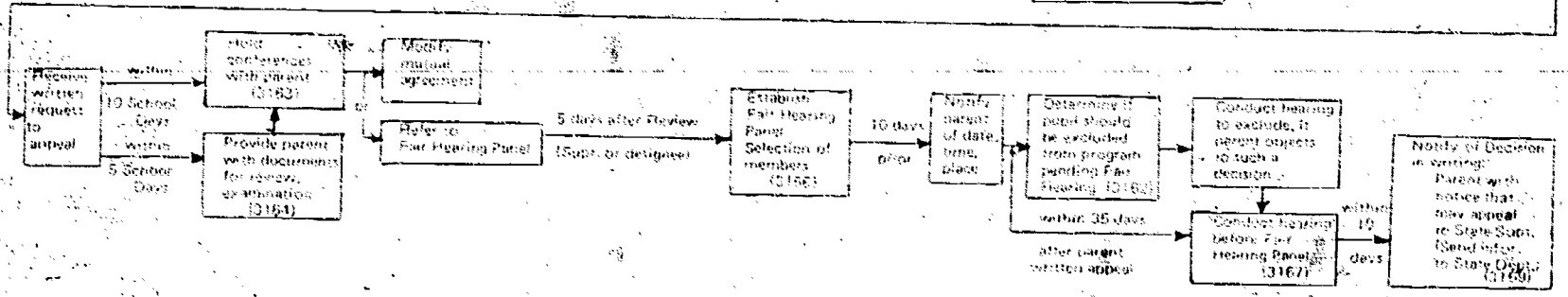
Office of the Santa Clara County Superintendent of Schools
San Jose, California
Glen W. Hoffman, Superintendent

PROCEDURAL DUE PROCESS SAFEGUARDS

ASSESSMENT TIMELINES:



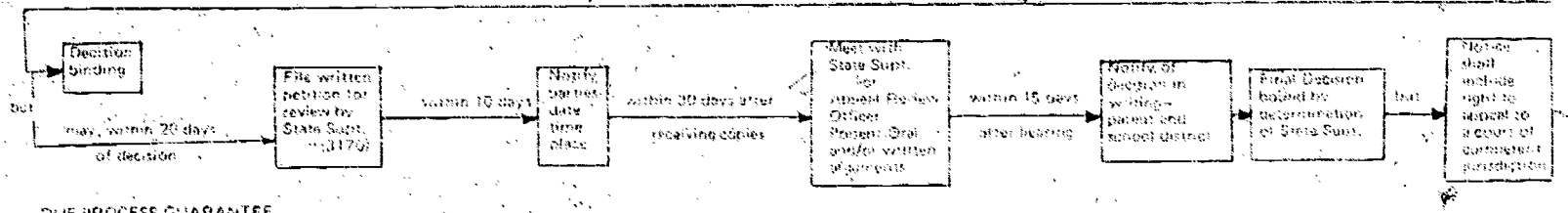
FAIR HEARING TIMELINES



(continued on next page)

PROCEDURAL DUE PROCESS SAFEGUARDS

FURTHER APPEAL TIMELINES:



DUE PROCESS GUARANTEE
PARENT OR LOCAL EDUCATIONAL AGENCY

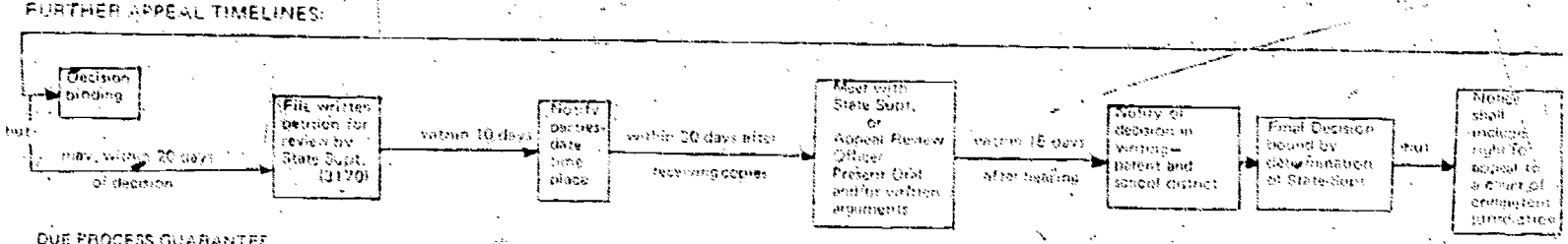
may request hearing on action re

1. pupil's identification as IVEN
2. pupil's assessment and implementation of IEP
3. denial, placement, transfer or termination of the pupil in a special education or related services program
4. pupil record data

Fair Hearing completed within 45 days of receiving complaint, or may agree on extension (3160).

Source: Office of the Santa Clara County Superintendent of Schools. Procedural due process safeguards. San Jose, California: Office of the Santa Clara County Superintendent of Schools, no date.

PROCEDURAL DUE PROCESS SAFEGUARDS



DUE PROCESS GUARANTEE
PARENT OR LOCAL EDUCATIONAL AGENCY

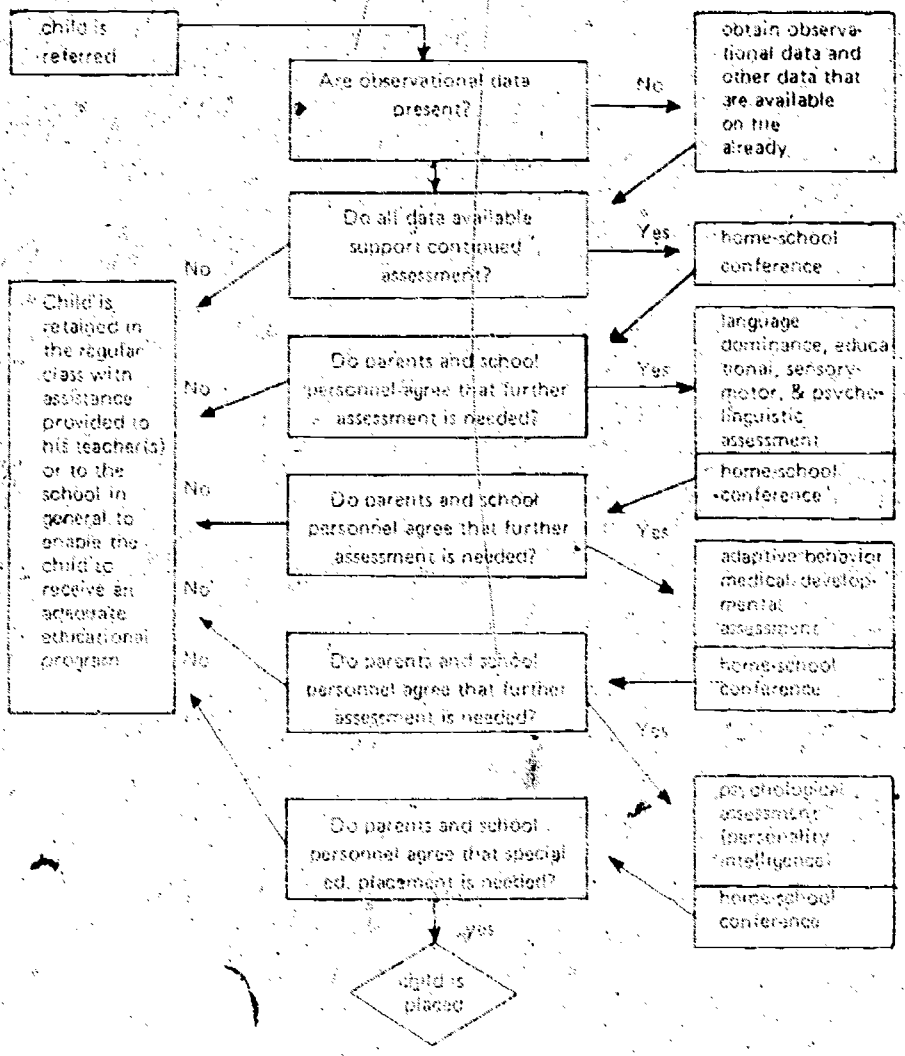
may request hearing on action re:

1. child's identification as IEP
2. child's assessment and implementation of IEP
3. denial, placement, transfer or termination of the child in a special education or related services program
4. [illegible]

Fair Hearing completed within 45 days of receiving complaint, SP may agree on extension (3160).

Source: Office of the Santa Clara County Superintendent of Schools. Procedural due process safeguards. San Jose California: Office of the Santa Clara County Superintendent of Schools, no date.

COMPREHENSIVE INDIVIDUAL ASSESSMENT
For Possible Mildly Handicapping Conditions



Source: Tucker, J. A. Operationalizing the diagnostic-intervention process. In Coordinating Office for Regional Resource Centers (Editor). *With bias toward none*. Lexington, Kentucky: Coordinating Office for Regional Resource Centers, 1976.

REFERENCES

Abeson, A., Bolick, N., and Hass, J. *A primer on due process.* Reston, Virginia: The Council for Exceptional Children, 1975.

American Association of Psychiatric Services for Children, Inc. *Developmental review in the early and periodic screening, diagnosis and treatment program.* Washington: U. S. Department of Health, Education, and Welfare, Health Care Financing Administration, The Medicaid Bureau. April, 1977.

Bloom, B. S. (Editor). *Taxonomy of educational objectives. Handbook I: Cognitive domain.* New York: McKay, 1956.

Brinegar, L. Partners in Learning: Focus of the California master plan for special education. In J. Jordan (Editor), *Teacher, please don't close the door.* Reston, Virginia: The Council for Exceptional Children, 1976.

Carroll, A., Gurski, G., Hinsdale, K., and McIntyre, K. *Culturally appropriate assessment. A source book for practitioners.* Los Angeles: California Regional Resource Center, 1977.

Cronbach, L. J. *Essentials of psychological testing.* New York: Harper and Row, 1970.

Davis, F. B. (Editor). *Standards for educational and psychological tests.* Washington: American Psychological Association, 1974.

De Avilia, E. Mainstreaming ethnically and linguistically different children: An exercise in paradox or a new approach? In Jones, R. L. (Editor), *Mainstreaming and the minority child.* Reston, Virginia: The Council for Exceptional Children, 1976.

Dent, H. E. Assessing black children for mainstream placement. In Jones, R. L. (Editor) *Mainstreaming and the minority child.* Reston, Virginia: The Council for Exceptional Children, 1976.

Fenton, K. S. et al. Role expectation: Implications for multidisciplinary pupil programming. Washington: U. S. Office of Education, Bureau of Education for the Handicapped, Division of Innovation and Development, State Programs Studies Branch. No date.

Fenton, K. S., Yoshida, R. K., Maxwell, J. P., and Kaufman, M. J. A decision model for special educational planning teams. Washington: U. S. Office of Education, Bureau of Education for the Handicapped, Division of Innovation and Development, State Programs Studies Branch. No date (a).

Fenton, K. S., Yoshida, R. K., Maxwell, J. P., and Kaufman, M. J. Recognition of team goals: An essential step toward rational decision making. Washington: U.S. Office of Education, Bureau of Education for the Handicapped, Division of Innovation and Development, State Programs Studies Branch. No date (b).

Goslin, D. A. *Teachers and testing*. New York: Russell Sage Foundation, 1967.

Harrison, D. B. The resource teacher. In J. Jordan (Editor) *Teacher, please don't close the door*. Reston, Virginia: The Council for Exceptional Children, 1976.

Hively, W. and Reynolds, M. C. (Editors). *Domain-referenced testing in special education*. Reston, Virginia: The Council for Exceptional Children, 1975.

Hoepfner, R., Strickland, G., Stangel, G., Jansen, P. and Patalino, M. *CSE elementary test evaluations*. Los Angeles: Center for the Study of Evaluation, UCLA, 1970.

Jones, R. L. Accountability in special education: Some Problems. *Exceptional Children*, 1973, *39*, 631-643.

Jones, R. L. Evaluating mainstreaming program impact on minority group children. In Jones, R. L. (Editor), *Mainstreaming and the minority child*. Reston, Virginia: The Council for Exceptional Children, 1976 (a).

Jones, R. L. *Standardized group and individual testing in the San Francisco Unified School District: Report and Analysis*. San Francisco: San Francisco Public Schools Commission, 1975 (b).

Jones, R. L. Special education and the future: Some questions to be answered and answers to be questioned. In forthcoming volume on the future of special education being edited by Maynard Reynolds, Leadership Training Institute/Special Education, University of Minnesota. In press, 1978.

Jones, R. L. Gottlieb, J., Guskin, S., and Yoshida, R. Evaluating mainstreaming programs: Models, caveats, considerations, and guidelines. *Exceptional children*. In press, 1978.

Jones, R. L. and Wilderson, F. Mainstreaming and the minority child: An overview of issues and a perspective. In Jones, R. L. (Editor), *Mainstreaming and the minority child*. Reston, Virginia: The Council for Exceptional Children, 1976.

Kirp, D. L. and Kirk, L. M. The legislation of the school psychologists' world. *Journal of School Psychology*, 1976, *14*, 83-89.

Krathwohl, D. R., Bloom, B. S., & Masia, B. *Taxonomy of Objectives: The classification of educational goals. Handbook II: Affective domain*. New York: McKay, 1964.

Lambert, N. M. and Hartsough, C. S. *The stress of school project* Berkeley: School of Education, University of California, 1971.

Lambert, N. M., Hartsough, C. S., Caffrey, C. M. and Urbanski, C. *Lexicon for apple observations*. Berkeley: School of Education, University of California, 1976.

MacMillan, D. L., Jones, R. L., and Meyers, C. E. Mainstreaming the mildly retarded: Some questions, cautions, and guidelines. *Mental Retardation*, 1976, *14*, 3-10.

MacMillan, D. L. and Meyers, C. E. The nondiscriminatory testing provision of 94-142. *Viewpoints*, 1977, *53*, 39-56.

Mid-East Regional Resource Center and National Association of State Directors of Special Education. *Child Identification: A handbook for implementation*. Washington, D. C. Mid-East Regional Resource Center, 1976.

Morrisey, P. A. and Safer, N. Implications for special education — The individualized education program. *Viewpoints* 1977, *53*, 31-38.

National Association of State Directors of Special Education. *Functions of the placement committee in special education*. Washington: National Association of State Directors of Special Education, 1976 (a).

National Association of State Directors of Special Education. *The Prince William Model. A planning guide for the development and implementation of full services for all handicapped children*. Washington: National Association of State Directors of Special Education, 1976 (b).

Office of the Santa Clara County Superintendent of Schools. Procedural due process safeguards. San Jose, California: Office of the Santa Clara County Superintendent of Schools, no date.

Orasanu, J., McDermott, R., and Boykin, W. A critique of test standardization. *Social policy*, 1977, *8*, 61-67.

Sabatino, D. (Editor). *Learning disabilities handbook: A technical guide to program development.* DeKalb, Illinois: Northern Illinois University Press, 1976.

Samuda, R. J. Problems and issues in the assessment of minority group children. In Jones, R. L. (Editor), *Mainstreaming and the minority child.* Reston, Virginia: The Council for Exceptional Children, 1976.

Sattler, J. M. *Assessment of children's intelligence.* Philadelphia: W. B. Saunders Company, 1974.

Semmel, M. I., Semmel, D. S., and Morrissey, P. A. *Competency-based teachers education in special education: A review of research and training programs.* Bloomington, Indiana: Center for Innovation in Teaching the Handicapped, Indiana University, 1976.

Subcommittee on the Handicapped of the Committee on Labor and Public Welfare, United States Senate. *Education of the handicapped act.* Washington: U. S. Government Printing Office, 1976.

Tucker, J. A. Operationalizing the diagnostic/intervention process. In Coordinating Office for Regional Resource Centers (Editor). *With bias toward none.* Lexington, Kentucky: Coordinating Office for Regional Resource Centers, 1976.

Weirdberg, R. A. and Wood, F. H. (Editors). *Observations of pupils and teachers in mainstream and special education settings: Alternative Strategies.* Minneapolis: Leadership Training Institute/Special Education, 1975.

Weiss, C. H. Interviewing in evaluation research. In Struening, E. L. and Guttentag, M. (Editors). *Handbook of evaluation research, Volume 1.* Beverly Hills: Sage Publications, 1975.

White, J. Toward a black psychology. In Jones, R. L. (Editor). *Black psychology.* New York: Harper and Row, 1972.

Yoshida, R. K. and Gottlieb, J. A model of parental participation in the pupil planning process. *Mental Retardation*, 1977, *15*, 17-20.

Yoshida, R. K., Fenton, K. S. and Kaufman, M. J. Evaluation of Education for the handicapped. *Phi Delta Kappan*, 1977, pp. 60-61.

Yoshida, R. K., Fenton, K. S., Maxwell, J. P., and Kaufman, M. J. Group decision making in the planning team process: Myth or reality? Washington: U. S. Office of Education, Bureau of Education for the Handicapped, Division of Innovation and Development, State Program Studies Branch. Research Report No. 2. No date (a).

Yoshida, R. K., Fenton, K. S., Maxwell, J. P. and Kaufman, M. J. Ripple effect: Communication of planning team decisions to program implementers. Washington: U. S. Office of Education, Bureau of Education for the Handicapped, Division of Innovation and Development, State Program Studies Branch. Research Report, No. 3, No date (b).

Yoshida, R. K., Fenton, K. S., Maxwell, J. P., and Kaufman, M. J. Parental involvement in the special education pupil planning process: The school's perspective. Washington: Bureau of Education for the Handicapped, Division of Innovation and Development, State Program Studies Branch. No date (c).

SECTION II

Protection in Evaluation Procedures

Jane R. Mercer

MERCER, JANE R. Dr. Mercer is a Full Professor in the Department of Sociology, University of California, Riverside. She received her Ph.D. in Sociology from the University of Southern California, 1962. Working as a Research Specialist for the California Department of Mental Hygiene, she directed an eight year study of the epidemiology of mental retardation funded by the National Institute of Mental Health. As an outgrowth of this earlier study, Dr. Mercer investigated the feasibility of a System of Multicultural Pluralistic Assessment (SOMPA) by studying representative samples of 2,100 black, Hispanic, and white elementary school students in California public schools. Her other research investigations studied the types of educational processes in desegregated schools which lead to equal status outcomes and positive mental health.

INTRODUCTION

The purpose of this paper is to develop criteria for implementing the provisions of Public Law 94-142 as described in the statute, in the congressional reports relating to the statute, and as further developed in the Final Regulations published in the *Federal Register*, Tuesday, August 23, 1977, Part II. The focus of this paper is on those aspects of the statute which deal with equal protection of the laws in testing materials and procedures used for the purpose of evaluating and placing handicapped children.

The paper will be divided into four sections. Chapter 1 will summarize the major components of Public Law 94-142, the congressional reports, and the final regulations which are to serve as benchmarks for the development of criteria. Chapter 2 will present the assessment models which are implied by the major components of the law. It will discuss the assumptions, characteristics, and limitations of each model. It will conclude with the definitions of crucial terms, such as "validity," "bias," and "fairness," which are appropriate for each model. Chapter 3 will present a design for an assessment procedure which incorporates all three models. If properly implemented, it can yield "racially and culturally nondiscriminatory" outcomes. Chapter 4 will present a series of checklists and ratings which a governmental or educational agency could use to evaluate the extent to which a particular set of procedures fulfills the requirements of the prototype procedures.

CHAPTER 1: REVIEW OF MAJOR LEGAL REQUIREMENTS FOR PROTECTION IN EVALUATION PROCEDURES

Types of Evaluations

According to the definition in the *Federal Register*, "evaluation means procedures used in accordance with 121a.530-121a.534 to determine whether a child is handicapped and the nature and extent of the special education and related services that the child needs. The term means procedures used selectively with an individual child and does not include basic tests administered to or procedures used with all children in a school, grade, or class" (subpart E, Procedural Safeguards 121a.500, definitions of "consent," "evaluation," and "personally identifiable"[c]).

Two types of evaluations are included in the stipulations concerning protection in assessment: the preplacement evaluation and the reevaluation.

1. *Preplacement evaluation.* The preplacement evaluation occurs before any

87

action is taken relating to placement of a handicapped child in a special education program. "Before any action is taken with respect to the initial placement of a handicapped child in a special education program, a full and individual evaluation of the child's educational needs must be conducted in accordance with the requirements of 121a.532" (20 U.S.C. 1412[5] [C]).

2. *Reevaluation.* In addition, the protection in assessment requirements also apply to the procedures to be used in the triennial reevaluations mandated by the law. "Each state and local educational agency shall insure. . . . (b) that an evaluation of the child, based on procedures which meet the requirements under 121a.532, is conducted every three years or more frequently if conditions warrant or if the child's parent or teacher requests an evaluation (20 U.S.C. 1412[5] [C]).

Consequently, the approaches suggested in this paper apply to both the preplacement evaluation and subsequent reevaluations which will be made of those children assigned to special education programs.

Purposes of the Evaluation Procedures

The Senate Report on the statute (Senate Report No. 94-168, Education for All Handicapped Children Act, June 2, 1975, pp. 28-29) and the rules and regulations published in the *Federal Register* both imply that there are two major purposes for conducting a preplacement evaluation or a reevaluation. One purpose is to identify the nature of the handicapping condition so that appropriate services can be made available. The second purpose is to provide detailed information on a student's current educational functioning so that an intervention program can be developed tailored to the needs of that student.

1. *To Identify the Nature of the Handicapping Condition.* The following sentence in Senate Report No. 94-168 describes this purpose. "In the educational process, the appropriate identification of handicapping conditions must take place in order to assure that a child receives appropriate services designed to meet his or her needs."

2. *To Assess Educational Need.* The following statements describe the second purpose for evaluation. "Such identification must also take place in order that a state or local educational agency may plan for the provision of appropriate services to meet the child's unique needs" (Senate Report No. 94-168). The rules and regulations specify that "tests and other evaluation materials include those tailored to assess specific areas of educational need and not merely those which are designed to provide a single general intelligence quotient" (*Federal Register* 121a.532).

88

The two purposes are related to the two dimensions of tests identified by Carver (1974): the psychometric and the edumetric. They are also conceptually separable into the "diagnostic" function and the "intervention" function. The former concentrates on (a) ascertaining the historical and etiological sources of the handicapping condition and (b) describing the current characteristics of the individual. The latter focuses on prescribing appropriate interventions. Thus, procedures designed for preplacement evaluation or reevaluation under Public Law 94-142 should include some measures directed at the identification of the nature of the handicapping condition and some measures directed at assessing educational needs.

The Nature of "Nondiscriminatory" Assessment

Although the term "nondiscriminatory" is not defined in the statute, the Senate Report, or the *Federal Register*, there are several statements which provide some insight into the intent of the law. The term is used in two different contexts to apply to two different populations: children with physical handicaps and children from minority racial and cultural groups.

1. *Nondiscriminatory Assessment of Children with Physical Handicaps.* The rules and regulations introduce the concept of "nondiscriminatory" assessment as avoiding "erroneous" classifications of children with physical handicaps (Federal Register 121a.532 [C]). The statements concerning such evaluations present the clearest definitions of the meaning of "nondiscriminatory" evaluation available in the government papers. "Tests are selected and administered so as best to ensure that when a test is administered to a child with impaired sensory, manual, or speaking skills, the test results accurately reflect the child's aptitude or achievement level or whatever other factors the test purports to measure, rather than reflecting the child's impaired sensory, manual, or speaking skills (except where those skills are the factors which the test purports to measure)."

2. *Racially and Culturally Nondiscriminatory Assessment.* The language of the statute and ancillary documents indicates that the law is also concerned with the "erroneous classification" of children variously described as "non-English-speaking," "poor," "minority," and "bilingual." The statue states that each state shall establish "procedures to assure that testing and evaluation materials and procedures utilized for the purposes of evaluation and placement of handicapped children will be selected and administered so as not to be racially or culturally discriminatory" (P.L. 94-142, 612[5] [C]). Mr. Miller is reported to have said: "A major problem which has been a concern of Congress and others is the manner by which children are identified as handicapped. Serious charges have been made, and substantiated, which indicate that some of the testing used to uncover children with 'learning disabilities' or more specific handicaps is discriminatory. They have been found biased against culturally deprived and

non-English-speaking children. I cite as a specific example the very sound work of Dr. Robert L. Williams of Washington University in St. Louis, who recently showed that some black youth perform poorly on some intelligence tests because of the vocabulary used. Congress last year attempted to remove such biases by writing into section 613 provisions banning discriminatory testing" (*Congressional Record-House*, July 29, 1975, Mr. Miller, p. H7763).

Further insight is provided by Senate Report No. 94-168 which speaks repeatedly of the "misclassification" of children. "The Committee is deeply concerned about practices and procedures which result in classifying children as having handicapping conditions when, in fact, they do not have such conditions." A major issue "with respect to problems of identification and classification" is "misuse of identification procedures or methods which results in erroneous classification of a child as having a handicapping condition." It is clear that the Committee is primarily concerned with taking "further steps in this legislation to provide that positive action be taken against erroneous classification of poor, minority, and bilingual children and against the invalid use of testing."

When the statement from the *Federal Register* (121a.532) which deals with the erroneous classification of children with physical handicaps is rephrased as follows, the statement provides a definition of "nondiscriminatory" assessment which can be applied as well to racially and culturally nondiscriminatory testing. Underlined phrases are those which have been reworded from the original. "Tests are selected and administered so as best to ensure that when a test is administered to a child *from a cultural background markedly different from the culture of the school* the test results accurately reflect the child's aptitude or achievement level or whatever other factors the test purports to measure, rather than reflecting the child's *cultural background* (except where *that background* is the factor which the test purports to measure)."

Multidimensional Assessment

The statute states that "no single procedure shall be the sole criterion for determining an appropriate educational program for a child" (P.L. 94-142, 612[5], Eligibility). Senate Report No. 94-168 directs the Commissioner to issue regulations to assure that "no single test or type of test or procedure is used as the sole criterion for placement and that all relevant information with regard to the functional abilities of the child is utilized in the placement determination."

The rules and regulations in the *Federal Register* are even more specific about the variety of tests and procedures which should be included. "The child is assessed in all areas related to the suspected disability, including, where

90

appropriate, health, vision, hearing, social and emotional status, general intelligence, academic performance, communicative status, and motor abilities" (121a.532[3][F]). In the "comment" following the above statement, reference is made to "psychological, physical, or adaptive behavior" assessments. Placement procedures are to "draw upon information from a variety of sources, including aptitude and achievement tests, teacher recommendations, physical condition, social or cultural background, and adaptive behavior" (121a.533[A][1]).

The "comment" following the above material makes it clear, however, that not all dimensions need be measured in every instance. "The agency would not have to use all the sources in every instance. The point of the requirement is to insure that more than one source is used in interpreting evaluation data and in making placement decisions. For example, while all of the named sources would have to be used for a child whose suspected disability is mental retardation, they would not be necessary for certain other handicapped children, such as a child who has a severe articulation disorder as his primary handicap."

Characteristics of Measures

Two characteristics of the measurement instruments are specifically mentioned in the statute and/or ancillary documents: test validation and test language.

1. *Validation.* The rules and regulations published in the *Federal Register* specify that tests and other evaluation materials shall "have been validated for the specific purpose for which they are used." The term "validation," however, is left undefined. The procedures for validating any particular measure for the specific purpose for which it is used are not specified. Hence, the concept of validation must be defined and operationalized in establishing a set of procedures to provide equal protection of the law in evaluation procedures.

2. *Native Language.* The statute also specifies that a child must be tested in his or her native language, if it is feasible. "Such materials or procedures shall be provided and administered in the child's native language or mode of communication, unless it clearly is not feasible to do so." However, it does not specify the conditions under which testing a child in his or her native language is required nor the procedures which are to be used in developing, administering, and interpreting such measures.

Characteristics of Placement Personnel

Senate Report No. 94-168 indicates a clear concern that there has been misuse and misinterpretation of assessment data in the past. A major concern "with

91

respect to problems of identification and classification is . . . the misuse of appropriate identification and classification data within the educational process itself." Two general proposals are made to ameliorate this situation: training of personnel and the use of the placement conference.

1. *Training of Assessment Personnel.* The rules and regulations stipulate that tests and evaluation materials be administered "by trained personnel in conformance with the instructions provided by their producer."

2. *Multidisciplinary Placement Conferences.* The legislation and the rules and regulations make the assumption that wider representation in the placement-planning conference is likely to reduce misclassification. For example, Senate Report No. 94-168 states: "The Committee has designed the individualized planning conferences as one method to prevent labeling or misclassification." In the *Federal Register* the rules and regulations stipulate that placement procedures should "insure that the placement decision is made by a group of persons, including persons knowledgeable about the child, the meaning of the evaluation data, and the placement options." They do not go so far as to recommend that a parent or a community advocate be included in the placement conference. However, the possibility of including nonschool persons is implied in the phrase "persons knowledgeable about the child." Related to this requirement is the provision that "the evaluation is made by a multidisciplinary team or group of persons, including at least one teacher or other specialist with knowledge in the area of suspected disability." No mention is made of including the child in the placement conference.

Summary of Major Legal Components

Analysis of the legislation, the congressional reports, and the Final Regulations in the *Federal Register* covering protection in evaluation procedures indicates that the following components will be needed in the theoretical models and operations proposed to evaluate the extent to which educational agencies are meeting the requirements of the law.

1. Preplacement *and* reevaluation procedures are to be included in the evaluation.

2. Evaluation procedures are to serve two purposes: identifying the nature of the handicapping condition and assessing educational needs.

3. Nondiscriminatory assessment is concerned both with children who have physically handicapping conditions and children who are "poor," "minorities," and/or "bilingual."

4. Assessment is to be multidimensional. It is to cover "health, vision, hearing, social and emotional status, general intelligence, academic performance, communicative status, motor abilities" and "psychological, physical, and adaptive behavior" assessments. Sources of information are to include "aptitude and achievement tests, teacher recommendations, physical condition, social or cultural background, and adaptive behavior."

5. Measures are to be validated for the specific purpose for which they are used and are to be provided in the child's native language or mode of communication.

6. Assessment procedures are to be conducted by trained personnel acting as a multidisciplinary team.

CHAPTER 2: ASSESSMENT MODELS UNDERLYING THE REQUIREMENTS OF P. L. 94-142

To meet the mandate for multidimensional assessment, which includes assessment of organic anomalies, adaptive behavior, and general intelligence, requires the evaluator to use three assessment models: the medical model, the social adaptivity model, and the general intelligence model. Each of these assessment models answers a different set of questions about the child, is based on its own definition of normal/abnormal, its own set of assumptions, and its own set of values. Each model has a distinct set of characteristics which distinguish it from the other models. These characteristics influence the types of measures appropriate for each model. Central to the task of this paper is the fact that each model has a different approach to the issue of determining the validity of a measure and each generates a markedly different definition of the nature of test "bias," "fairness," and "racially and culturally nondiscriminatory assessment." Proper interpretation of measures based on each model requires that those involved in making an assessment have an understanding of the assumptions and limitations of the model(s) which they are using as well as training in the administration of specific tests.

Chapter 2 of this paper will outline, briefly, the three assessment models which evaluators will be using if they are to meet the mandate of Public Law 94-142 (Mercer, in press).

The Medical Model

The medical model is the most familiar of the three models. It has been called the pathological model, the deficit model, the disease model, and the clinical model. It is the conceptual model developed in medical research to understand

and combat pathological conditions in the organism.

Questions Addressed by the Model.

The medical model is designed to answer questions about the state of the organism. Hence, the measures of "health, vision, hearing . . . and motor abilities" (*Federal Register* 121a.532 [3] [F]) and of "physical condition" (*Federal Register* 121a.533[A][1]) mandated by Public Law 94-142 will require evaluators to utilize the medical assessment model.

Definition of Normal/Abnormal.

An abnormality in the organism is defined as a process that tends to destroy the biological integrity of the organism as a living system and to interfere with its functioning. Such pathological processes are identified by their biological symptoms. In the medical model, normal tends to be a residual category which consists of those persons who do not manifest the symptoms of pathology. In some situations, such as spastic or palsied conditions, behavioral patterns are interpreted as symptoms of organic malfunctioning. In such cases, there is the assumption that the behavior is the result of pathological conditions in the organism and not the result of learning.

Assumptions of the Medical Model.

The medical model assumes that the symptoms are caused by some biological condition in the organism — a disease process, lesion, chromosomal anomaly, or other pathological condition. When the only observable symptoms are behaviors and the organic basis for the condition cannot be specifically identified, the burden of proof rests with the evaluator who uses the medical model to present evidence that organic inferences are justified.

A second assumption of the medical model is that the sociocultural characteristics of the individual are irrelevant to making a diagnosis or prescribing a treatment. The physician does not need to know the language which a person speaks or the cultural heritage in which a patient has been reared to diagnose and treat tuberculosis, cancer, measles, near-sightedness, a hearing loss, or other pathologies. The *sine qua non* of the medical model is evidence of an organic basis for the condition.

Universal Value Frame.

The medical model is based on a universal set of values derived from the fact that the human organism is similar in all human societies. There is a single, cross-cultural set of norms which can be applied to assess the health status of any human being, regardless of cultural setting. These norms are based on the nature of the human organism and do not vary with language or culture. Consequently, the medical model is not culture bound. The human organism responds in similar fashion to physical trauma and disease processes regardless of the cultural milieu.

Focus of Assessment.

In the medical model, the organism is the focus of assessment and pathology is perceived as a condition existing in the person, an attribute of the organism. Thus, we say a person *is* tubercular or *has* scarlet fever. It also holds that a pathological condition can exist, unrecognized. Hence, within this model, it is logical to "screen" populations for undiagnosed or undetected pathologies.

Properties of Measurement Instruments.

Although the medical model recognizes degrees of pathology, for the most part it is dichotomous. A person either has the measles or does not have the measles, has tuberculosis or does not have tuberculosis, and so forth. The model tends to divide persons into those who are "negative," i.e., have no signs of a given pathology and those who are "positive," i.e., have signs of a given pathology. For this reason, many medical model measures have low ceiling and have distributions of scores which are negatively skewed.

On the other hand, the degree of deficit is carefully measured. For example, once a person achieves 20/20 vision, no attempt is made to determine if vision is better than 20/20. Once a person has "passed" an auditory acuity measure, no attempt is made to detect higher levels of acuity. Medical model measures tend to focus on deficits, counting the number of errors.

Validity.

The validity of a medical model measure is determined by the extent to which it predicts pathology. The validity of the urine test for phenylketonuria is determined by its ability to predict children who will later develop the symptoms of PKU if not placed on a special diet. The validity of a blood test for syphilis is determined by its ability to predict who will develop symptoms of advanced syphilis if they go untreated.

Operationally, validity of medical model measures is determined by intercorrelating scores from a variety of medical model measures i.e. health history, physical dexterity, sensory-motor coordination, medical history, etc. Correlations should be statistically reliable but we would not expect them to be large. Conversely, correlations with sociocultural characteristics should be low *unless* there is a clearly established genetic or other biological link between such characteristics and pathological symptoms. See the earlier discussion in this section.

Racial and Cultural Discrimination.

Because the medical model measures the intactness of the organism, its norms are biologically determined and universal to the species. Its norms are not culture bound. They are not determined by human decision. Diagnosis and treatment are not culture bound. It is not necessary to take the sociocultural background of the person into account when interpreting his or her performance. For example, there is a single norm for visual acuity which can be used

95

equally appropriately in the United States, Russia, or southern India. For this reason, when measures are used which meet the assumptions of the medical model, the evaluator need not be concerned with issues of racial or cultural discrimination. The medical model is not sensitive to racial or cultural characteristics, per se.

The above statement does not mean that there is no correlation between racial and cultural characteristics and some biological pathologies. For example, sickle cell anemia is more common among Afro-Americans and Tay-Sachs disease is more common among Jews. It cannot be concluded from this correlation, of course, that there is something about the Afro-American culture which produces sickle cell anemia or that there is something about the Jewish religion which produces Tay-Sachs disease. Rather, biological causes are sought. The high prevalence for these conditions within each group is related to clearly identified genetic factors. There is evidence for biological transmission of these pathologies.

Diagnostic Values.

The pervasive code in medical decision-making holds that it is worse for a physician to overlook a pathology than it is for him to suspect pathology, continue to make diagnostic tests, and later to find that there is no pathology (Scheff, 1966). The belief that a "false negative" is a more serious error than a "false positive" is based on the presumption that an untreated pathology may worsen and eventually lead to death, while additional diagnostic tests will not be harmful to the patient. When an evaluator is using the medical model to screen for possible biological anomalies, using statistically standardized measures, the evaluator should apply the values of the medical ethic in setting cut-off levels for screening. A conservative level designed to avoid "false negatives" would be one standard deviation below the mean on standardized measures for which there are no medically determined norms.

The Social Adaptivity Model

The social adaptivity model has also been called the social system model (Mercer, 1973, in press), the social deviance model, and the social competence model (Doll, 1953). It is a conceptual model which focuses on the assessment of behavior in a social setting in relation to a set of social norms.

Questions Addressed by the Model.

The social adaptivity model is designed to answer questions about the *behavior* of an individual in a social setting as contrasted with the medical model which is designed to answer questions about the state of the organism. In general, tests of "social and emotional status," "academic performance," "communicative status," and adaptive behavior are measures of behavior in a social setting.

Achievement tests and teacher recommendation, which are also mentioned in the federal regulations for implementing Public Law 94-142, likewise qualify as measures of behavior. When evaluators are using instruments to assess behavior in a social setting, they are operating within the social adaptivity model.

Definition of Normal/Abnormal.

Each social system is composed of social statuses, social roles, and social norms. The social statuses are the positions which persons occupy in a system, such as the position of teacher, pupil, principal in the school. Associated with each status is a social role which consists of the behaviors of persons occupying the status. Persons participating in a social system share common expectations concerning the appropriate behavior for persons playing particular roles in the system. These shared behavioral expectations are the norms of the system. The term "normal" from a social deviance perspective does *not* relate to biological signs or have any necessary relationship with the biological organism. Normal behavior is that which conforms to the expectations of other members of the group. Deviant or abnormal behavior is that which does not meet group expectations.

Assumptions of the Social Adaptivity Model.

The first assumption of the social adaptivity model is that there are multiple definitions of "normal" behavior which vary from social system to social system and from role to role. It is "normal" for a student to yell and jump around when attending a pep rally, but such behavior would be judged as "abnormal" in an arithmetic class. There is nothing, per se, which makes yelling and jumping either "normal" or "abnormal." It is only in relation to the norms of a particular social setting that the normality of the behavior can be evaluated. Thus, the social adaptivity model is highly situation specific. To determine whether a given behavior is "normal," one must know (1) the social system in which the person is operating, (2) the status the person holds in the system and its associated role, (3) the role expectations or norms for judging the performance of persons playing that role, and (4) the behavior of the person to be evaluated.

A second assumption of the social adaptivity model is that the norms are not biologically determined. The norms by which a person's role performance is judged are evolved in a political-definitional process. Those persons having the greatest power in the social system impose their definition of "normal" on less powerful members. Typically, parents, who are the most powerful members of the family, impose their behavioral norms on their children. Teachers, the most powerful person in the classroom, typically impose their behavioral norms on their students. In the larger society, those cultural groups which are dominant politically and economically impose their behavioral standards on less politically powerful groups.

The normative system which currently guides American public education was

established as a result of the political and economic dominance of the Anglo-American cultural group. Public school norms require that all instruction be in English, that a student's language development be determined by evaluating proficiency in the English language, and so forth. The growth of the testing movement has been intimately associated with the public schools. Standardized academic achievement and aptitude tests are *indirect* measures of the extent to which a student's academic performance meets the expectations of the school. Teacher evaluations are *direct* measures of the extent to which a student's performance meets school norms. Both types of measures focus on evaluating behavior in the school setting and are examples of measures which fit the social adaptivity assessment model.

Social System Specific Value Frame.

Unlike the medical model, which is based on universal values, the social adaptivity model is based on the values of each social group. Those values are revealed through the norms governing each social role in each social system. Hence, the social adaptivity model is a multinormative model which has as many sets of norms as there are social systems and statuses and roles within social systems. The values of the school may differ from the values of the student's home or the values of the peer group. The behavior of a student may fulfill role expectations in one group but may violate the norms of another. Transsystem and transcultural interpretation of the "normalcy" of behavior are inappropriate. Even generalization beyond the specific role being evaluated cannot be justified unless the roles and normative structures of the two systems are closely linked. An example of closely linked normative structures would be high school to college or college to graduate school.

Focus of Assessment.

Unlike the medical model which views pathology as an attribute which the organism carries into every social situation, the social adaptivity model focuses on assessing behavior. Abnormality or normality are judgments about behavior, not about the state of the organism. Abnormality or normality is *not* regarded as a characteristic or trait of the individual whose behavior is being evaluated. Behavior is role specific and norms are role specific. Hence, judgments about behavior are inevitably tied to the norms governing the specific role behaviors being evaluated.

Properties of Measurement Instruments.

Because social groups recognize unacceptable, acceptable, and exemplary role performance, the social adaptivity model is both a deficit and an asset model. The full range of measurement is possible. Therefore, instruments designed as direct or indirect measures of role performance should have a full distribution of scores and tend to form a normal distribution. Items should reflect behaviors valued by the group and should have a high enough ceiling to identify outstanding performers.

Validity

The validity of a social adaptivity measure is determined by its ability to identify those persons who are succeeding and those who are failing to meet group expectations. The assessment of family role performance should reflect the family's expectations; the assessment of peer group performance should reflect peer group evaluations; the assessment of academic role performance should reflect teacher evaluations, and so forth. The predictive validity of academic achievement and aptitude tests has traditionally been measured by their ability to identify those students who are judged as competent or as incompetent by the teacher. Criterion-related validity, as currently defined by persons in psychological measurement, conforms to a social adaptivity assessment model definition of validity (Cleary et al., 1975).

It is important to note that the person designing a test of social adaptivity does not impose a set of values on the system within which behavior is being evaluated. System values are taken as a "given" when measuring social adaptivity, whether that system be the child's family, the school, the peer group, or the ethnic community. Social system measures should embody the viewpoint and the evaluations of persons in the systems in which the child is trying to achieve an adaptive fit. Traditional psychometric definitions of "criterion-related validity" fit the social adaptivity assessment model. The test designer accepts the value judgments of persons in the system as to the types of behaviors which are "socially relevant and useful" or "socially important" (Cleary et al., 1975, p. 23). "Criterion related validity. . . is simply the extent to which test scores are related to a socially important criterion measure" (Cleary et al., 1975, p. 25). The decision as to which behaviors are "socially important" is determined by the relative power relations in the social system.

It should be noted, that this definition requires a *direct* measure of social role performance as the criterion i.e. school grades, teacher rating, peer rating, etc. Two *indirect* measures i.e. test-test correlations such as "intelligence" tests correlated with "achievement" tests are not appropriate measures of the validity of a social adaptivity assessment.

Another issue is that of determining the extent to which a particular *direct* measure of a child's performance in a particular social system accurately represents the evaluations of other members of the social system. Does a particular teacher's rating of a child's performance in school reflect the rating which would be given by other teachers? Does the peer rating of one member of the peer group adequately capture the ratings of other peer group members? Such questions refer to the issue of inter-rater reliability. They are answered by correlating the judgements of various members of a social system with each other. For example, correlations between the reports of a child's adaptive behavior secured independently from a child's mother and father would provide information on inter-rater reliability in the family social system. Correlations

between ratings given by various teachers would provide information on the reliability of a particular teacher's assessment of the child's adaptive fit in the school, and so forth. The judgements of persons who are non-members of the social system *cannot* be used to determine the reliability of the assessments made by system members. Non-members are outside the normative structure of the group. As outsiders, they are not privy to the subtleties of the behavioral norms which operate in the group. Knowledge of those norms is what differentiates the "insiders" from the "strangers". For example, a mother's report of a child's adaptive fit in the family cannot be validated against a report of a social worker or a teacher. They are "outsiders". Their judgements reflect the social norms of the systems in which they operate and not that of the family.

Racial and Cultural Discrimination.

When measuring a child's adaptive fit to a particular social system, the question of racial and cultural discrimination is directly related to the accuracy of the test in predicting the evaluations which system members will make of the child's performance. Traditional psychometric definitions of test "fairness" fit the social adaptivity model. Cleary et al. (1975) have presented a clear statement of the nature of test "fairness" when operating from a social system assessment model. "A test is considered fair for a particular use if the inference drawn from the test score is made with the smallest feasible random error and if there is no constant error in the inference as a function of membership in a particular group" (Cleary et al., 1975, p. 25). Operationally, a test would be considered "fair" or "unbiased" if the following four conditions are met. (1) The variances are homogeneous for the populations being compared. (2) The correlation coefficients between the test scores and group evaluations of individual role performances are nonzero and equal for persons of different racial and ethnic groups. (3) The regression lines for persons of different racial and ethnic groups are parallel, i.e., have similar slopes. (4) The regression lines for persons of different racial and ethnic groups have similar intercepts.

In a social adaptivity assessment model, the fact that one group may have a higher average score on a test than another group or that one group may receive higher average ratings on their social role performance in a particular social system than another does not mean that the social adaptivity measure is racially or culturally discriminatory, so long as the four conditions listed above are met. For example, if an evaluator were interested in predicting which students would perform successfully in the social systems of the inner city, he might ask them to take a test to measure their knowledge of black English on the assumption that such knowledge would be required for acceptable role performance in the inner city. The Black Intelligence Test for Cultural Homogeneity (BITCH-100) developed by Robert Williams (1975), a vocabulary test of 100 words selected from the *Dictionary of Afro-American Slang*, could be used as the predictor. The

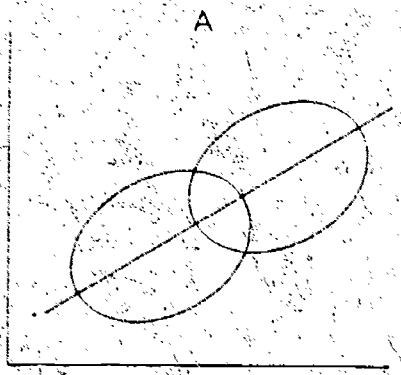
fact that students from a rural southern background or from a middle class background might earn lower average scores on the BITCH-100 than students reared in the central city would not be evidence that the test was racially or culturally biased, if the test predicted accurately which students would perform their social roles in the most acceptable fashion as evaluated by persons of the inner city. Likewise, the fact that central city youth might be rated higher, overall, in their performance by other persons of the inner city would not be considered evidence of "bias." If the correlation coefficients were of similar magnitude, the regression lines were parallel, and the intercepts were the same, the fact that the average scores for rural or for middle class students on the test and on the criterion might be lower than for central city students would not be regarded as evidence of racial or cultural discrimination within the social adaptivity assessment model. Such a situation is depicted in Figure 1-A.

However, if one or more of the four criteria are not met, then the test is defined as "biased" and "unfair." A typical situation in social adaptivity measures is depicted in Figure 1-B. In this case, the average scores of one group on both the test and the criterion are lower than the scores of the other group. The regression lines are parallel but have different intercepts. When the regression line of the higher scoring group is used to predict the performance of the lower scoring group on the criterion measure, the performance of the lower scoring group is overpredicted. That is, the lower scoring group is predicted to perform better in their social roles in the social system than they are actually likely to perform. If the two groups are combined to calculate a joint regression line, the joint line will fall between the lines for the individual groups. Prediction from the joint regression line will likewise overpredict the lower scoring group's probable social role performance. Hence, this situation discriminates against the higher scoring group and in favor of the lower scoring group.

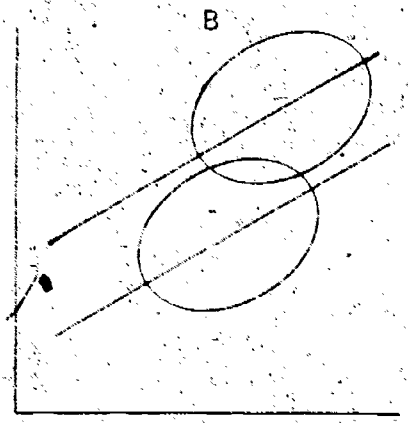
A situation in which a test meets none of the criteria for a "nonbiased" measure is depicted in Figure 1-C. If the intercepts for the regression lines are different and the lines are not parallel, the regression lines will cross. Accurate prediction of the criterion performance is not possible. When the majority regression line is used to predict minority role performance, minority performance is overpredicted above the point at which the lines cross but underpredicted below the point at which the lines cross. Below the point of crossing is precisely the portion of the distribution in which [redacted] decisions are made about special education placement. If the low scoring group is a racial or cultural minority, as is frequently the case with standardized "intelligence" tests, and the regression lines cross, the situation would lead to underestimating the actual student role performance of low scoring minority children.

To summarize, the traditional psychometric definition of test "bias" fits the social adaptivity model. The definition takes the values of the social system as a

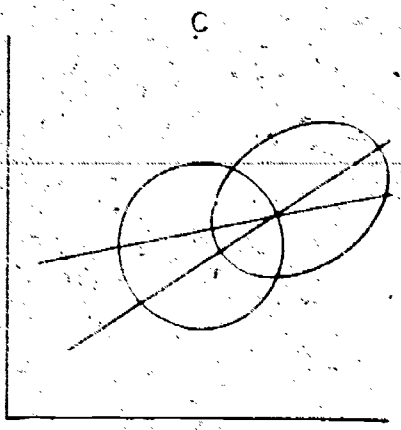
FIGURE 1



An "unbiased" measure: two groups with the same correlation coefficients, same slopes for regression line, and same intercept for regression line.



A "biased" measure: two groups with the same correlation coefficients and regression slopes but different intercepts; overpredicts group with lower average scores.



A "biased" measure: two groups with different correlation coefficients, regression line slopes, and regression line intercepts; overpredicts above the corssed lines and underpredicts below the crossed lines.

"given" and focuses attention on the accuracy of a measure in predicting role performance, as measured by some criterion, such as ratings by members of the group. If the prediction is made with the smallest feasible error and there is no constant error as a function of racial or cultural group, then the test would be regarded as an unbiased measure. To determine whether a test is racially or culturally discriminatory when predicting social role performance within the social adaptivity assessment model, the user must know whether the correlation coefficients are similar, whether the regression lines are parallel, and whether the regression lines have the same intercepts for members of different racial and cultural groups. For a more detailed explanation of the definition of racial and cultural discrimination which is appropriate to the social adaptivity assessment model, see Cleary et al. (1975).

Purposes for Using the Social Adaptivity or Social System Model.

There are two distinct but related purposes for assessing a child's behavior within the social adaptivity or social system assessment model. (1) One purpose for assessing behavior is to identify the nature of "handicapping condition" so that programs, treatments, services, and other resources can be made available to the student. Identification of the "handicapping condition" is required by statutes, such as Public Law 94-142, which make funding of educational services for children contingent on labeling the child as having a particular "disability." Such categorical aid programs require classification before treatments and services can be made available to the pupil. Thus, this function is closely related to administrative decisions. (2) A second purpose for assessing behavior is to provide information on the pupil's educational needs in order to develop educational interventions, preferably interventions designed as part of a coherent individual educational plan. The two purposes are similar to the two dimensions of tests described by Carver (1974): the psychometric function and the edumetric function.

The psychometric function of tests focuses on providing measures which will produce a relatively stable ranking of individuals on the behavior measured. The aim of psychometric tests is to highlight differences among individuals and to rank them according to their proficiency in performing some type of socially valued role behavior. For this reason, scores on psychometric tests are ordinarily standardized relative to some normative population. Standard scores identify the individual's relative position in the distribution of scores of persons on whom the test was normed. Little change in relative position is anticipated from one administration of the test to the next, since scores are ordinarily standardized by age. Such stability in relative rank is considered an indication that the test is "reliable."

The edumetric function of tests focuses on providing information on the student's current performance. This information is to be used to design

103

educational interventions in specific academic areas. Edumetric tests are closely tied to the educational curriculum. The pupil's performance is evaluated against a series of graduated educational objectives or criteria which have been established as representing stages in accomplishing the goals of the curriculum. Ordinarily, raw scores rather than standardized scores are used. It is anticipated that a pupil's performance as measured by the raw scores will improve with instruction. Slight emphasis is placed on comparing the pupil's performance with that of other pupils. Rather, the individual pupil's performance is measured repeatedly and is compared over time. The expectation is that a student's performance will progress through the graduated series of educational objectives, if given appropriate instruction.

Both psychometric and edumetric tests fit within the social adaptivity or social system model because they both measure learned behavior which is evaluated against standards which have been set by a social group. In general, tests which have been designed for psychometric purposes, such as so-called tests of intelligence and aptitude, are not adequate as edumetric measures. For example, no one would attempt to design an educational program to teach a child arithmetic on the basis of arithmetic subtest scores on the WISC-R. On the other hand, edumetric measures are not generally useful for psychometric functions.

Diagnostic Values.

Diagnostic values within a social system assessment model differ markedly from those used when operating within a medical model. The rationale for these values is based upon the sociological concept of primary and secondary deviance and the deviance process.

There are four stages in the development of a deviant career. Stage 1, primary deviance, occurs when the behavior of an individual is first labeled as abnormal by someone in the system. Stage 2 emerges when others in the system take action to counteract the behavior labeled as deviant. Negative sanctions may be applied in an attempt to "normalize" the behavior by bringing it into conformity with system norms. Stage 3 is reached when there is role reorganization within the sytem. The person exhibiting behavior defined as deviant is moved from the status of "normal" into a status reserved for those who cannot or will not fulfill "normal" role expectations. Large social institutions, such as the schools, have formalized many deviant statuses to accommodate persons with a large variety of deviant behaviors, such as educable mental retardate, learning disabled, educationally handicapped, and so forth. In more extreme cases, the offending member may be removed from the social system by being expelled from any status in the system. Stage 4, secondary deviance, occurs when the individual defined as a deviant internalizes the deviant role, restructures the self in terms of the deviant status, acts out the deviant expectations, and accepts the deviant definition of the self as appropriate.

Labeling behavior as deviant initiates the deviance process. Because social systmes are powerful agencies which can mold and shape the careers of participants and, ultimately, may influence the individual's social and psychological identity, erroneous labeling of behavior as deviant may have serious negative consequences for the individual launched on a deviant career. Thus, within the social adaptivity model, the ethical code is to avoid labeling any behavior as deviant if there is a shadow of a doubt about the reliability or validity of the label. Conversely, the ethic of the social adaptivity assessment model would support maintaining a child in the status of "normal" as long as possible.

Because psychometric tests are used for making administrative decisions in categorizing children for programs and are based on norms which directly compare a child's performance with that of other children, the interpretation and use of such tests is likely to initiate the deviance process. For this reason, a low cut-off level for defining a person as "subnormal" or behaviorally deviant is recommended. In general, two or more standard deviations below the mean for the standardization sample, or the lowest 2.5 percent, provides such a conservative criterion (Mercer, 1973, chapter 14). This criterion conforms with the current definition of the American Association for Mental Deficiency which defines subnormal performance as scores more than two standard deviations below the mean on a standard measure (Grossman, 1973).

The problem of initiating the deviance process is less acute when using edumetric measures, such as teacher constructed tests and tests linked to evaluating the achievement of specific objectives within a criterion-referenced curriculum. Since the child's present performance is compared with his or her own past performance rather than with a normative population, the testing focuses on designing interventions rather than making comparisons with others. However, it should be recognized that invidious comparisons are still possible even when using edumetric tests if children who are further along the curricular continuum are compared with children who are at earlier stages in the curricular continuum.

The "General Intelligence" Model

The general intelligence model, which has also been called the pluralistic model (Mercer, in press), was first introduced by Alfred Binet when he developed a test of learning from which he attempted to make inferences concerning the "intelligence" of the child. The attempt to make inferences about a child's intelligence, aptitude, potential, or mental ability from test performance has been one of the major threads in testing for the past seventy-five years.

The *Federal Register* lists measures of "general intelligence" among those tests which may be appropriate in the assessment of some types of "disability" (Rules

105

and Regulations, 121a.532, Evaluation Procedures [F]). It requires that "tests and other evaluation materials include those tailored to assess specific areas of educational need and not merely those which are designed to provide a single general intelligence quotient" (Rules and Regulations, 121a.532, Evaluation Procedures [B]). The implication of this statement is that assessment procedures will include some tests "designed to provide a single general intelligence quotient," but that such measures cannot stand alone.

Questions Addressed by the General Intelligence Model.

The general intelligence assessment model attempts to answer questions about the individual's mental ability, intelligence, or potential for learning. It views "intelligence" as an attribute which the individual can apply in coping with solving problems in new situations. For example, Terman and Merrill (1960, p. 5) identify one of the distinctive characteristics of the Binet-type scale as "the concept of the measurement of a 'general intelligence' which functions as mental adaptability to new problems." Wechsler (1974, pp. 3-7) provides the following definition:

> Intelligence is the overall capacity of an individual to understand and cope with the world around him. . . (1) It [the definition] conceives of intelligence as an overall or global entity; that is, a multidetermined and multifaceted entity rather than an independent, uniquely-defined trait. (2) It avoids singling out any ability, however esteemed (e.g., abstract reasoning), as crucial or overwhelmingly important. . . Ultimately, intelligence is not a kind of ability at all, certainly not in the same sense that reasoning, memory, verbal fluency, etc., are so regarded. Rather, *it is something that is inferred* from the way these abilities are manifested under different conditions and circumstances. One can *infer* an individual's intelligence from how he thinks, talks, moves, almost from any of the many ways he reacts to stimuli of one kind or another. Indeed, historically, appraisal of such responses has been the usual way of judging intelligence. . . [Inferences are made by] comparing each subject's test performance not with a composite age group but exclusively with the scores earned by individuals in a single (that is, his or her own) age group. . . Each person tested is assigned an IQ which, at his age represents his relative intelligence rating. This IQ, and all others similarly obtained, are deviation IQs since they indicate the amount by which a subject deviates above or below the average performance of individuals of his own age (emphasis added).

These are three aspects of Wechsler's discussion which are central to the general intelligence model. Intelligence is (1) conceived as a global entity or capacity, that is (2) inferred from the current behavior of the individual (not measured directly), by (3) comparing the individual's performance with others of the same age. The inferential model is clearly statistical. It defines intelligence in terms of the relative rank of the person's performance compared with others of the same age. Indeed, Wechsler is quite forthright in declaring that "no attempt has been made to define a priori the social and clinical significance of any given IQ." In short, no argument is made for the predictive validity of the test in the sense in which "validity" is defined within the social system model. (See the earlier discussion of the social system model.)

106

Finally, Wechsler recognizes that a person's performance and, hence, the inferences which can be made from that performance, are influenced by cultural and socioeconomic background. He emphasized the importance of "the examiner's awareness of the degree to which a subject's responses may be influenced or conditioned by his cultural and socioeconomic background" (Wechsler, p. 7) but provides no procedure for estimating that influence or taking it into account in interpreting test scores.

Definition of Normal/Abnormal

The definition of "normal" and "abnormal" within the general intelligence model is essentially a statistical definition. The statistical definition of "normal" is familiar to anyone who has been introduced to the concept of the normal curve. In this model, an individual is described by his or her relative position in a frequency distribution of scores of other persons who have taken the same test. The "norm" for the test is the statistical average for the population on whom the test was standardized. The general intelligence model defines abnormality according to the extent to which an individual varies from the average of the population on a particular set of behaviors.

Establishing the statistically normal is a straightforward process. The investigator specifies the population of persons on which the norms will be based and then measures the entire population or a representative sample of the population on the behaviors being normed. Scores on the measure are organized into a frequency distribution, and the average score — i.e., the statistical mean — is calculated. The mean is accepted as the norm. Customarily, persons with scores that deviate not more than one standard deviation above or below the mean are regarded as falling in the "normal range" and make up approximately 68 percent of the population. Therefore, in the statistical definition, normal equals the statistical mean plus or minus one standard deviation from the mean.

In establishing a statistical norm, the test maker uses the characteristics of the particular population being studied to establish the boundaries of "normal." When the population on which the test is normed is changed, the boundaries of "normal" will also be modified. In a fundamental sense, the persons constructing and standardizing a particular test determine that range of behaviors which will be considered "normal" when they made the decision concerning the population on which the test shall be normed (Mercer, 1973, pp. 2 ff.).

There is a long tradition in psychological testing supporting the importance of developing "local" norms for tests in those situations in which a local population cannot be considered to be part of the universe from which the sample was selected for purposes of norming the test. When groups can be shown to be from different populations, statistically, it is not appropriate to combine those

populations for purposes of calculating a single norm for the test (Mercer, 1972). When a statistical model is used to define normal performance, the norms emerging from measurements taken on one population cannot be safely generalized beyond that population. Unlike the medical model, statistical definitions of "normal" are neither transsocietal nor universal. They are tied to the population on which the test was normed.

Assumptions of the Model

In the "general intelligence" model, inferences are made about the individual's "intelligence" based on his or her relative position in the distribution of scores of other persons on whom the test was normed. Obviously, intellectual capacity cannot be measured directly, because that would require assessment of the genetic component of performance, the genotype. An individual's genotype can only be expressed through behavior learned in a social and cultural setting, his phenotype. The test measures what a person has learned, his phenotype. On the basis of his performance, inferences are made about the nature of his genotype. The general intelligence model assumes that it is possible to make valid inferences about the genotype from a properly normed and administered test. Persons using the "general intelligence" model are constantly making inferences about genotypes on the basis of the performance of phenotypes. The logic behind these inferences is relatively simple, but the assumptions are rarely met in actual practice.

(1) If two persons have an equal opportunity to learn the types of cognitive, linguistic, and mathematical skills and to acquire the types of information in the test, (2) if they are equally motivated to learn these skills and types of information, (3) if they are equally motivated to exert themselves in a test situation and equally familiar with the demands of the test situation, (4) if they are equally free of emotional disturbance and anxieties that might interfere with their performance, (5) if they are equally free of biological dysfunctions and organic difficulties that might have interfered with their learning the materials in the test or might interfere with test performance, *then* any difference between their performance on a test that measures the extent to which they have learned these cognitive, linguistic, and mathematical skills and acquired certain types of knowledge is probably the result of differences in their intellectual endowment or learning potential. Simply stated, if learning opportunities and all other factors are equal, those persons who learn the most and who perform the best probably have the greater mental capacity than those who learn the least and perform most poorly. Of course, the major difficulty in applying this logic in interpreting test scores within a general intelligence model is that all factors are seldom equal.

Value Frame for the Model.

The "general intelligence" model is based on the premise that persons who learn quickly and who can solve problems effectively are valued in all human societies and that such potential should be cultivated, both for the benefit of the individual and the benefit of the larger society. Therefore, it is important in the assessment process to identify learning potential in children which may not have been recognized because of physical handicaps or sociocultural differences between the background of the child and the expectations of the school.

Focus of Assessment

The focus of assessment is on the behavior of the child in the test situation. However, the evaluator is concerned with going beyond a description of the test behavior and making inferences about the student's general intelligence. Cautious inferences about a child's probable learning potential may be made from his or her performance on a test, *if* the assumptions of the model can be met.

Properties of Measurement Instruments.

Instruments typically used in the general intelligence model are so-called tests of intelligence, such as the various Wechsler scales and the Stanford-Binet. Such tests were first developed by Binet and Simon in the early 1900s to identify those French children who would not be likely to benefit from a regular public school education. Binet attempted to choose items for his test with which all persons participating in French society would be familiar. This practice has continued. Consequently, the abilities and skills and knowledge covered in the typical test are selected from the particular cultural stream which is dominant in a particular society (Mercer, 1973). Although some items, such as arithmetic, may be directly related to academic curricula, others are drawn from the general cultural pool of the dominant cultural group (Mercer and Brown, 1973). The language of the test is the language of the dominant cultural group. The tests attempt to measure how much the individual has learned about the language, style of thought, history, political and social institutions of the dominant culture and the extent to which he or she has acquired the cognitive skills valued by the dominant group. Any test of an individual's learning i.e. achievement in a particular cultural setting could be used as the basis for inferring learning potential, provided the assumptions of the inferential model for inferring "intelligence" are met. However, in individual assessment, the individually administered test is preferred because there are fewer uncontrolled variables, such as the child's reading ability, the child's motivation to pursue a paper and pencil task, and so forth.

The interchangeability of so-called tests of "intelligence" and tests of "achievement" is cogently described by Wesman (1968), Jencks (1972), and Cleary et al. (1975). Wesman states, "All ability tests — intelligence, aptitude, and achievement — measure what the individual has learned — and they often

109

measure with similar content and similar process . . . Such justification as we have for our labeling system resides entirely in the purpose for which the test is used, not in the test document itself. If our intent is to discover how much the examinee has learned in a particular area, . . . We label the test an 'achievement' test. If our intent is to predict what success an individual is likely to attain in learning a new language, or a new job, we seek those specific previous learnings the possession of which bodes favorably for that future learning, and we label the test an 'aptitude' test or a 'special' aptitude test. If our intent is to predict future acquisition of learning over broad areas of environmental exposure, we seek those previous learnings the possession of which will be relevant to as many future learning situations as we can anticipate. This test we label an "intelligence" test. The selection of test items or sample tasks for the three purposes may or may not differ; but in each instance what is measured is what was previously learned. We are not measuring different abilities; we are merely attending to different criteria." (Wesman, 1968, p. 269.)

Jencks states, "Many (test manufacturers) distinguish, for example between 'achievement' and 'aptitude'. In principle, achievement tests tell whether students have mastered some body of material that the tester deems important. Aptitude tests theoretically tell whether students are *capable* of mastering a body of material the tester deems important. In practice, however, all tests measure *both* aptitude *and* achievement. . . If two students have had the same opportunity to acquire verbal skills, and if one has picked them up while the other has not, the test does indeed measure 'aptitude'. But if one child has been raised speaking Spanish and another English, the test measures the Spanish-speaking child's mastery of a foreign language. If the Spanish-speaking child does worse than the English-speaking, this shows lower achievement in this area, but it need not imply less aptitude. . . When everyone is equally well prepared, achievement tests become aptitude tests. When people are unequally prepared, aptitude tests become achievement tests. In light of this, we will not make any rigid distinction between aptitude and achievement. We will simply try to use the term that seems appropriate in a given context." (Jencks, 1972, pp. 55-56) When persons undertaking the assessment of children use the three assessment models correctly, they differentiate carefully between the use of a test as a measure of "achievement" within the social adaptivity model and as a measure of "intelligence" within the pluralistic or general intelligence model.

Finally, Cleary et al (1975) reach a similar conclusion. "There are no differences in kind, as noted earlier, between intelligence and achievement, or between aptitude and achievement. There are, instead, four dimensions appropriate to the description of tests and the repertoires they sample." The four dimensions discussed are breadth, the extent to which a test is defined by a specific educational program, the recency of the learning sampled, and the purpose of the test. The authors then continue, "The dimensional analysis is useful in indicating why there is confusion concerning the proper category in which to

110

place certain tests. Just because differences among test items are quantitative and not qualitative, it is possible for one man's intelligence test to be another man's achievement test. Thus, Jensen (1968) categorized the National Merit Scholarship Examination as an intelligence test, but precisely the same items were used in the Iowas Tests of Educational Development for assessing achievement." (p. 21).

Definition of Test Validity

Precisely because "one man's intelligence test can be another man's achievement test", it is the assessment model within which a test is used rather than the test per se which determines whether the use of the test for a particular purpose is "valid". The general intelligence model assumes that "intelligence" is an attribute of the person. It cannot be measured directly because current functioning reflects some combination of the person's biological endowment and the person's cultural exposure to the materials in the test. Therefore, the individual's "intelligence" has to be inferred using the inferential model described earlier in this paper. This model is essentially statistical. The validity of a particular inference within the general intelligence model is determined by the extent to which the procedures used to make an inference about the "intelligence" or "learning potential" of a particular person meet the requirements of the inferential paradigm on which such conclusions are based. If the person is being compared with others who have had the same opportunity to learn the materials in the test, have been similarly motivated to learn those materials, have had similar test taking experience, are equal with respect to emotional disturbance, anxiety, and physical disability, then the procedures meet the assumptions of the inferential model and are "valid". If any or all of the above conditions are not equal, then the procedures do not meet the assumptions of the inferential model and are "invalid." Thus, the requirements of the inferential model for estimating "learning potential" encompass the issues raised in P. L. 94-142 concerning both the nondiscriminatory assessment of children with physical handicaps and children from non-Anglo core culture backgrounds.

Validity within the general intelligence model does not relate to predicting some criterion performance. Criterion related validity is appropriate within the social adaptivity model. Wechsler alludes to this distinction when he declares that "no attempt has been made to define a priori the social and clinical significance of any given IQ"; rather, intelligence "is something that is inferred from the way these abilities are manifested under different conditions and circumstances. One can infer an individual's intelligence from how he thinks, talks, moves, almost from any of the many ways he reacts to stimuli of one kind or another." (Wechsler, 1974, pp. 3 - 7). In actual practice, it has been the logic of the inferential paradigm rather than some type of criterion related validity which has operated in clinical assessments. It is significant that during the testimony in Larry P. v Wilson Riles only *one* study could be found in which an individually

administered test of "intelligence" had been "validated" for elementary school children against a *direct* measure of social system performance such as grades earned in academic subjects or teacher ratings of academic performance. Apparently, the validity of using a test for the purpose of inferring "intelligence" has, in fact, rested on the logic of the inferential paradigm rather than empirical evidence of predictive validity.

Racial and Cultural Discrimination.

Charges that tests are racially and culturally discriminatory have been focused primarily on the interpretation of tests within the general intelligence model. (Bernal, 1975; Jackson, 1975; Williams, 1975). The definition of a racially and culturally discriminatory test within the general intelligence model differs markedly from that within the social adaptivity model. It relates to the assumptions of the inferential paradigm which must be met before judgements can legitimately be made about a child's "intelligence." Fundamentally, the argument is that children from those racial and cultural groups which do not fully share in the dominant Anglo core culture of American society do not have the same opportunities to learn the materials in the tests, the same motivation to learn the materials in the tests, nor the same test-taking experience as the children from the core culture with whom they are being compared. Therefore, no inferences can be made about their "intelligence" by comparing their performance on the tests with that of children who have had greater exposure to the culture from which the test items are drawn. Since the assumptions of the inferential model are not met, the tests are "biased" for making inferences about the mental capacity, "intelligence," or learning potential of non-Anglo children.

Thus, the definition of test bias in the general intelligence model corresponds closely to the dictionary definition of the term "bias," which means to show partiality, to favor unfairly, or to make inequitable comparisons (Webster's, 1966). In the general intelligence model, a test is biased if it shows partiality to one group by including mainly questions from their cultural heritage and few, if any, questions from the cultural heritage of other groups and then makes inequitable inferences about the "intelligence" of the groups on the basis of their responses to the questions.

Five major lines of evidence are used to establish the cultural bias of a test in this model.

(1) *Examination of test items.* If an examination of test items reveals that the questions, the test language, and the performances expected of children represent a single cultural heritage, this fact is taken as evidence that the test is biased (Mercer, 1975; Jackson, 1975; Williams, 1975).

(2) *Differences in Average Scores.* Differences in the average scores of different racial and cultural groups on the test are further evidence within the general

intelligence model that the test is biased. Differences which are large and statistically significant indicate that the two groups come from different populations and that the groups (a) should not be combined for purposes of establishing a joint-norm and (b) the norms established on one group are not appropriate for making inferences about the "intelligence" of the other group (Jackson, 1975).

It should be noted that in the general intelligence model differences in average scores on the test are cited as evidence of test bias. In the social adaptivity model, the psychometric definition of bias which is appropriate to the model states that differences in average scores are irrelevant to the question of bias so long as the predictions made from the test are unbiased. The social adaptivity model is concerned with bias in predicting future performance, while the general intelligence model is concerned with bias in making inferences about general intelligence. The purposes of the two models are different, and the definitions of what constitutes a racially and culturally discriminatory test is different for each model.

(3) *Heterogeneity of Minority Populations.* Minority groups in the United States are internally very heterogeneous. Some members of minority groups are structurally and culturally integrated into the Anglo-American core culture. Other members of minority groups are bicultural, participating partially in the Anglo core culture and partially in their non-Anglo tradition. Still other members of minority groups are completely outside the Anglo core culture. They may be recent migrants to the United States or members of groups, such as some native American tribes, who have insulated themselves from the dominant Anglo culture. In either case, children reared in such non-Anglicized settings may not speak the English language and may know little or nothing about the American core culture. It can be demonstrated that those minority children who are reared in families which are structurally and culturally integrated into the core culture perform better on tests of "general intelligence," such as the WISC-R, than children who are from bicultural backgrounds. Children from bicultural backgrounds perform better on such tests than those from families who are completely outside the Anglo core culture. Differences are systematic and linear. The average scores increase progressively as the children in a particular group are from more Anglicized backgrounds. The range of mean scores is approximately 15 points (Mercer, 1973). The finding that the average test scores on the Wechsler scales for minority children increase progressively as the families of the children are more acculturated to the Anglo core culture is presented as one line of evidence that the tests are sensitive to cultural differences and systematically discriminate against children from non-Anglicized backgrounds. Therefore, it is argued that such tests cannot be used to make inferences about general intelligence when testing minority children.

(4) *Experimental Studies.* A fourth line of evidence for cultural bias, when

113

making inferences within the general intelligence model, comes from studies which demonstrate that minority infants who are involved with their mothers, in socialization programs which increase their exposure to the language and practices of the dominant Anglo core culture perform significantly better on "intelligence" tests than comparable infants who have not been involved in such interventions. Differences as large as 33 points have been reported (Garber, 1975).

(5) *Minority Children Adopted into Core Culture Homes.* Findings that minority children adopted into middle class Anglo homes do significantly better than their counterparts who were not exposed to the dominant culture through adoption provide a fifth line of evidence for cultural bias in making inferences concerning general intelligence from test performance (Scarr and Weinberg, 1976).

Proposals for Eliminating Racial and Cultural Discrimination within the General Intelligence Model.

A complete discussion of the variety of proposals which have been made to correct for the cultural bias in tests is beyond the scope of this paper. Two general approaches have been proposed: developing new tests and modifying existing tests.

Developing New Tests.

(1) The search for the "culture-free" test was based on the assumption that it is possible to develop a test consisting of items which are "free" of all cultural influence. Efforts in this direction have been unproductive since all learning takes place in a socio-cultural setting and all tests measure learned behavior. Wesman (1968) and Williams (1975) both agree that the search for the culture-free test is futile.

(2) The "culture-fair" test has been pursued along two lines: the common-culture approach and the balancing items approach. The common-culture approach assumes that there are tasks or problems which are common to all cultures and that a test can be developed using only such items. "To be equally fair to all persons, an intelligence test should present problems that are equally familiar or equally unfamiliar to all" (Eells, et al., 1951, p. 16). However, Davis and Eells were unsuccessful in achieving culture-fairness with the common-culture approach based on a series of "games" (Cronbach, 1975). Research on the culture fairness of Cattell's Culture-Fair Tests for Measuring Intelligence (Institute for Personality and Ability Testing, 1973) is limited and inconclusive. Some investigators have attempted to use Raven's Progressive Matrices (Raven, 1960) and the Goodenough Draw-A-Man Test (Harris, 1963) for cross-cultural assessment on the assumption that the tasks required in these measures are common to all cultures. However, differences in average group performance persist (Dennis, 1966; Irvine, 1966). The psycho-situational context of the

testing situation presents numerous complexities in communication which make it difficult to interpret performance cross culturally (Mehan, 1973; Roth, 1974; MacKay, 1974; and Bersoff, 1973). Other investigators have proposed developing culture-fair measures by including an equal number of items from each of the cultural traditions of persons taking the test. For example, a test designed for Hispanic, black, and white children would include a third of the items from the Hispanic culture, a third of the items from the black culture, and a third of the items from the Anglo core culture. Breland, et al. (1974) has questioned the cross-cultural stability of test items. In addition, the political feasibility of requiring children from the politically dominant group to respond to questions from a minority culture is questionable.

(3) The "culture-specific" test has also been proposed. The BITCH-100, developed by Williams (1975) is an example of a test designed for persons from a minority culture. The primary difficulty with this approach is the tremendous cost of producing a large variety of tests, one for each sub-group within each ethnic group. No single test would be appropriate for all black children because their cultural backgrounds range from the peasant background of the rural south to the middle-class professional family. No single test would be appropriate for all Hispanic children, and so forth.

(4) Some investigators have hypothesized that the developmental stages described by Piaget could provide a cross-cultural framework for assessment. De Avila and Havassy (1975), have developed test instruments for this purpose.

Modifying Existing Tests.

(1) The translation of existing tests into languages other than English has been attempted as a means of controlling for the inappropriateness of an English language test when assessing a non-English speaking child. However, direct translation of vocabulary items is frequently not possible. Translation also changes the difficulty level of items because those words which are used frequently in one language system and hence have a lower difficulty level may not be used as frequently in another language system. The content of the items remains culture specific. Smith (1974) studied fourteen different versions of the Binet, revised for nine countries from 1908 to 1960, and concluded that each Binet item had a different cultural loading which was bound to a particular time and locale. Studies of Spanish translations of the Wechsler scales report similar difficulties (Moran, 1962; Coyle, 1965). If a test is translated, then the item difficulties will need to be recalculated; the item content should be changed to reflect the cultural milieu of the new language system; and the test would need to be re-normed on the population for which it is to be used. In short, a translated test is essentially a new instrument and must be restructured and restandardized.

(2) Changing the procedures for administering the test is another modification

which has been used. Some examiners have varied the speed and power components of the test based on the assumption that, given adequate time, persons from differing cultural background will perform in a similar fashion. Schwarz (1963) contends that there are unpredictable complexities when the speed-power factor is varied in test administration. Changing the wording of questions, providing additional cues, changing the scoring procedures, and other modifications in the standardized administration procedures make it impossible to interpret the scores within any existing normative framework. A new norming of the test based on the modified procedures is required before scores can be interpreted.

(3) A test-train-retest paradigm has been proposed by Budoff (1972) using nonverbal reasoning tasks such as those in Raven's Progressive Matrices. He interprets the gain score resulting from teaching as a measure of the child's "learning potential." Although this is a promising technique, there are difficulties in standardizing the "training" phase of the assessment and it is a lengthy procedure requiring several contact hours.

(4) Developing sociocultural norms for each sociocultural group within an ethnic group is another approach to achieving non-discriminatory assessment using presently available measures (Mercer and Lewis, 1978). The System of Multicultural Pluralistic Assessment (SOMPA) provides a procedure in which the average score for a child's sociocultural group is calculated using four measures of the child's background: Urban Acculturation, Socioeconomic Status, Family Size, and Family Structure. The child's score is then compared to the norm for his or her sociocultural group to determine how high or low the child's performance is when compared with others who presumably have had similar opportunities to learn the materials in the test. The comparison is converted to a metric which has a mean of 100 and a standard deviation of 15 for more ready interpretation and is called "Estimated Learning Potential." The SOMPA approach has the advantage of being inexpensive (calculations take about 3 minutes), feasible (the multiple regression equations are available for Hispanic, black, and Anglo children and are relatively easy to develop for other groups), and do not require extensive re-training of existing personnel. Distributions of scores on Estimated Learning Potential are completely normalized for all groups so that each group has a mean of 100 and a standard deviation of 15 and approximately the same percentage of children in the two tails of the distribution. When this procedure is used, the test content, procedures, and scoring are not altered in any way. Only the interpretative framework, the norms, are varied to reflect the appropriate comparison group for each child.

Diagnostic Values within the General Intelligence Model.

The diagnostic values appropriate to the general intelligence model are similar to those governing the social adaptivity model. In each case, diagnostic labels can trigger either positive or negative institutional responses which, in turn, may

propel a child into esteemed or disesteemed statuses, as the case may be. Just as the deviance process tends to culminate in the psychic restructuring of the individual in terms of the deviant status and role, so institutional processes which move the child into esteemed statuses associated with enhanced opportunities tend to culminate in the psychic restructuring of the child in terms of the valued status and role. Hence, overestimating a child's learning potential or intelligence is a less serious error than underestimating a child's learning potential because the consequences of the former error are likely to be positive, while those resulting from the latter error are likely to be negative.

CHAPTER 3: DESIGN FOR RACIALLY AND CULTURALLY NONDISCRIMINATORY ASSESSMENT PROCEDURES

Having identified some of the unique characteristics of each of the three assessment models used by persons making assessments of children, the next task is to discuss how the three models can be integrated into an overall design for racially and culturally nondiscriminatory assessment. A set of concepts first introduced by Cromwell, Blashfield, and Strauss (1975) will be used as the framework for the following discussion. They presented a set of criteria for a logical classification system which emphasized linkages between etiology, diagnosis, treatment, and prognosis. Although their interest was primarily from a medical perspective, the framework is useful in conceptualizing the assessment process in public education.

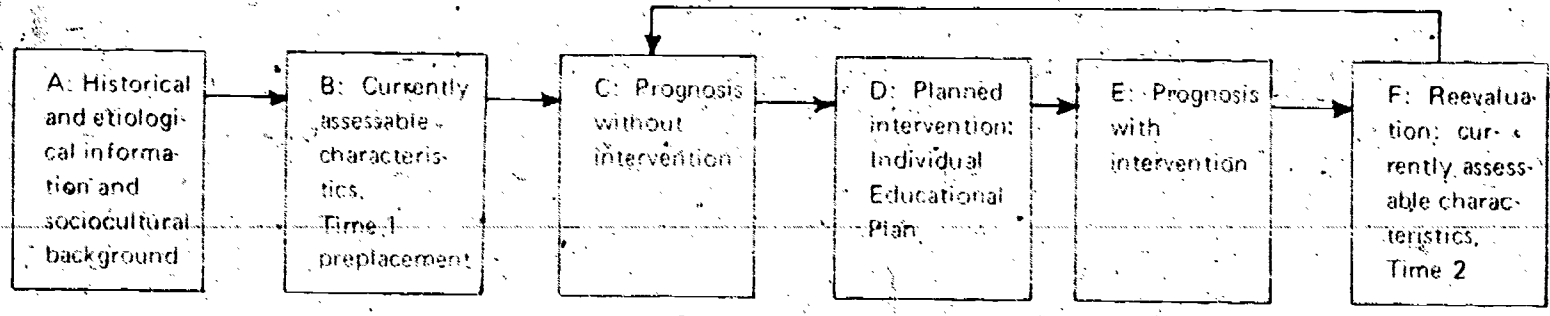
Building Diagnostic Constructs in Educational Assessment

Cromwell and his associates (1975) identified four major components in a complete diagnostic construct and gave them alphabetic designations. Two additional components have been added to include both the "preplacement" evaluation and the reevaluation. In the assessment process, a diagnostic construct is created by operationalizing each component. Figure 2 presents a pictorial representation of the building of diagnostic constructs.

A. *Historical and Etiological Information*. Component A consists of information about the child's developmental history, health history, and family background which may be useful in tracking possible sources of current difficulties. Ordinarily, the more complete the historical and etiological information, the more precise and comprehensive the understanding of the child's current characteristics. All three assessment models provide Type A information.

B. *Currently Assessable Characteristics* (Time 1). Component B consists of all

FIGURE 2
SIX COMPONENTS OF A DIAGNOSTIC CONSTRUCT



the information about the child's current characteristics which can be assessed and is relevant to developing an understanding of the situation at the time of the initial assessment, Time 1. Again, all three assessment models can provide Type B information.

C. *Prognosis without Intervention.* Component C has been added to the Cromwell (1975) framework. It consists of the prognostic statement based on the Type A and B information which presents the probable outcome if there is no intervention of any kind. If the conclusion is that the prognosis is favorable without intervention, then the building of the diagnostic construct would be terminated at this point in the assessment process. If, on the other hand, the prognostic conclusion is that outcomes will be negative if there is no intervention, the diagnostic process moves to D.

D. *Proposed Intervention: Individual Educational Plan.* Component D is the plan for the intervention which is developed on the basis of the information provided in components A and B. Proposed interventions will include treatments for identified organic problems, such as glasses to correct for visual impairment, hearing aids to correct for auditory impairment, physical therapy to ameliorate physical disabilities, and so forth. Interventions also include educational procedures to assist the child in mastering the curriculum of the school and to help the child achieve an adaptive fit in those systems in which the child may be having difficulty, such as the peer group, the community, the school, and so forth. The specific nature of these interventions is determined by the unique needs of each child as revealed in Components A and B and constitutes an Individual Educational Plan.

E. *Prognosis with Intervention.* Component E consists of the prognosis made by the assessment team of the probable outcome of the proposed interventions. This prognosis is developed as a series of specific objectives stated in terms of a series of time lines. The objectives include those relating to interventions directed at correcting or ameliorating biological problems, those relating to interventions directed at educational needs, and those relating to achieving an adaptive fit in the various social systems in which the child must operate. Thus, Component E provides the basis for monitoring the progress of the child in response to the proposed treatment by specifying precisely what objectives are to be met in each area of identified need by a particular time. The agreed-on time frame provides the time schedule for reevaluation to determine if the interventions are effective. The statement of objectives provides the array of assessments which will be made at the time of the reevaluation.

F. *Reevaluation: Currently Assessable Characteristics, Time 2.* Reevaluation takes place at the time agreed on in the Individual Educationl Plan. It covers the array of assessable characteristics which were incorporated in the Individual Educational PLan as targets for intervention. Based on the findings of

Component F, the diagnostic process recycles to Component C. Again, the assessment team estimates the probable outcome if there is no further intervention. If the decision is that the outcomes will probably be positive, then interventions may be terminated. If the decision is that the outcomes will probably be negative, then the process moves to Component D. An updated Individual Educational Plan is developed based on the child's assessable characteristeics at Time 2 and a new prognostic statement, Component E, is developed. The reevaluations and the recycling of the assessment procedures continue as long as the assessment team concludes that interventions are needed.

Partial Diagnostic Constructs

When the schema depicted in Figure 2 is used to evaluate the assessment process in a particular educational institution or school district, it can identify missing components in local assessment practices. Some partial diagnostic constructs are useful in spite of the fact that they are incomplete, while other partial diagnostic constructs are either useless to educators or invalid. The primary types of partial diagnostic constructs will be discussed briefly.

ACDE and BCDE constructs are useful even though they are based on incomplete information. The ACDE construct includes no information on the current characteristics of the child, and the BCDE construct includes no information on the history, etiology, or family background of the child. In some circumstances, it may be possible to arrive at a prognosis without intervention (C), develop an individual plan (D), and make a prognosis with intervention (E) on the basis of incomplete information, but such cases would be limited to those in which the child's problems were quite specific, such as a visual impairment correctable with glasses, a minor articulation problem, and so forth.

ABC, AC, BC constructs are those rare instances in which A and/or B information provides the basis for making a prognosis, but there are no known treatments for the child's problem. Hence, the diagnostic construct stops after the prognosis without intervention because there is no known intervention for the child's disabilties. Fortunately, with improved medical and educational technology, there are relatively few circumstances in which a child cannot be helped to some extent by some type of medical or educational intervention.

DE constructs are quite common in education. They occur when an intervention, such as a reading program or mathematics program, is instituted for a group of children, irrespective of their individual characteristics and without Type A or B information. A generalized prognosis is made that those receiving the program will reach some educational objective. Such constructs may be defensible in planning educational programs for children who have no special educational needs; however, they are not defensible in planning an educational program for "handicapped" children. Public Law 94-142 is quite specific on this point.

120

AB type constructs are useful to the scholar and academician who is trying to track relationships between historical or family background factors (Type A information) and the currently assessable characteristics of the child (Type B information). However, they are of no value to the educator because they are not linked to any prognostic statement or to any known intervention. For example, it may be interesting to know that children who come from large families (Type A information) tend to do less well on tests of academic achievement (Type B information), but this knowledge cannot be utilized directly by the school in planning for the individual child.

ABCD, ACD, BCD [redacted] are invalid. They include an intervention (the D component) for which there is no known outcome and hence no statement of the prognosis with intervention can be made. Any educational intervention, therapy, or program which is instituted without knowledge that it is efficacious would qualify as an invalid construct.

Building Diagnostic Constructs Using the Three Assessment Models

As noted earlier in this paper, Public Law 94-142 and the federal guidelines relating to that statute list a large number of measurements which are to be included in a multidimensional assessment. Some of these measures fit the medical assessment model, such as measures of health, vision, hearing, motor abilities, and physical condition. Other proposed measures fit the social adaptivity or social system model, such as measures of academic performance, teacher recommendations, "communicative status," and adaptive behavior. Still others relate to the measurement of "general intelligence," an inferential procedure which belongs in the general intelligence or pluralistic model.

Each of these models provides information which can be used at each stage of building a diagnostic construct for the individual child. Figure 3 presents a schematic representation of the major types of input provided by each of the assessment models at each stage in constructing a diagnostic construct.

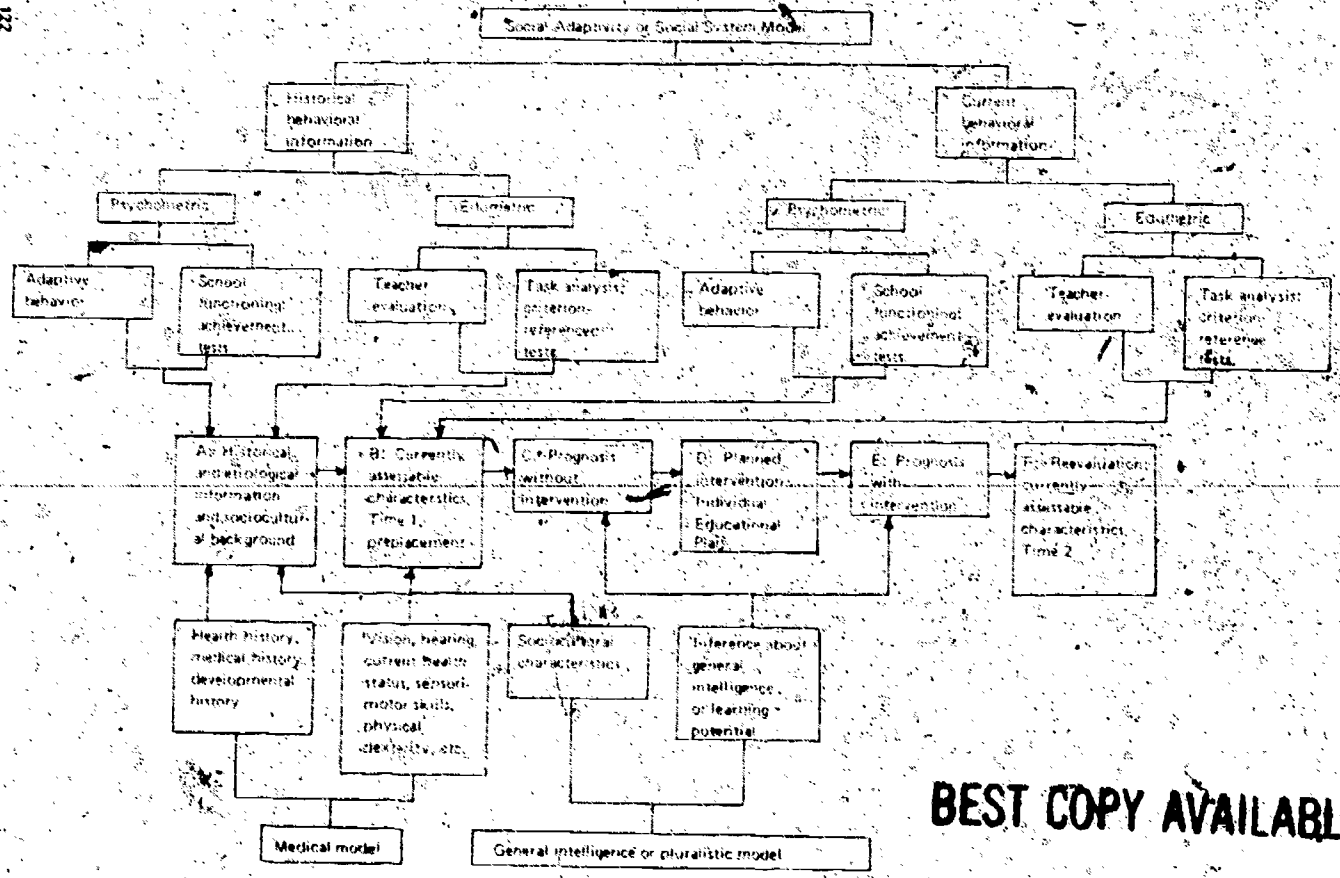
Medical Model Measures.

There are three types of measures within the medical model which provide Component A information: the health history, the medical history, and the developmental history. Ordinarily, information on the health history and the developmental history comes from the mother or other principal caretaker. The medical history will come from medical records kept by individual physicians, clinics, hospitals, or other health service institutions.

There are numerous types of medical model information which contribute to Component B. Tests of physical coordination or dexterity, tests of sensorimotor

121

FIGURE 3
ARTICULATION BETWEEN THE THREE ASSESSMENT MODELS
AND THE SIX COMPONENTS OF A DIAGNOSTIC CONSTRUCT



skill, tests of vision, tests of hearing, measures of height and weight, and the enumerable measures used in the assessment of a child's current health status by the physician all feed in crucial information on the current status of the biological organism. As noted earlier in this paper, the validity of such measures is determined by their ability to identify pathological conditions in the organism. When using such measures, either for screening or for a more comprehensive medical work-up, the practitioner need not be concerned with issues of racial and cultural discrimination because the medical model is a transsocietal model. Such measures can be administered and interpreted without reference to the racial and cultural background of the person being assessed. They are generally applicable to all members of the species, regardless of cultural heritage.

Social Adaptivity or Social System Model

Historical behavioral information (Component A) is frequently available for the child who has attended the school for a period of years. Such information becomes part of the assessment process by providing background information on the child's earlier performance on tests designed to measure behavioral variables. As mentioned earlier, there are two types of behavioral measures generally used in the schools: psychometric and edumetric. Figure 3 lists two types of psychometric measures mentioned in the guidelines for Public Law 94-142: measures of adaptive behavior and measures of academic achievement. In general, measures of adaptive behavior evaluate the child's adaptive fit in a variety of social systems: the family, the peer group, the community, the economy, the school, and so forth (Mercer, 1977). Measures of academic achievement focus exclusively on the child's skills in fulfilling the academic expectations of the student role in the public school.

In addition, the guidelines for Public Law 94-142 mention two types of edumetric measures: "teacher's recommendations" and measures of academic performance, presumably measures more closely related to the curriculum of the school than the typical standardized achievement test. There are numerous forms in which teachers can provide edumetric information about the child's behavior which can be useful in program planning: teachers' referral information, grades, checklist, ratings of student's performance, observations, and anecdotal information. A wide variety of edumetric tests has been developed during the past decade; but it is beyond the scope of this paper to discuss these in any detail. Such tests are used to assess the academic skill development of the child in specific academic areas. It is assumed that the development of more complex skills is contingent on mastery of lower-level skills. Such measures are subject-matter referenced. The specific skills or behaviors deemed to be important are determined by the goals, values, and objectives of the educational system, and skill development is influenced by sociocultural background. Hence, edumetric measures designed for "task analysis" belong within the framework of the general social system model (Mercer and Ysseldyke, 1976).

123

In the "preplacement" or initial evaluation of the child, historical information on the child's behavior as assessed on psychometric and edumetric measures may or may not be available. In either case, a full array of both psychometric and edumetric information should be gathered at the time of the evaluation.

The validity of such measures is determined by their ability to reflect accurately the degree to which the child is making an adaptive fit to the behavioral expectations of the role being evaluated. If the behavioral measure purports to assess the child's performance in the peer group, the validity of the measure is determined by the power of the score to accurately reflect the extent to which the child's performance is acceptable to the peer group. If the measure purports to evaluate the child's performance in the family, its validity is determined by the accuracy with which it captures the family's perception of the child's performance. If the measure purports to assess the child's adaptive fit to the role of student, its validity is determined by how well scores on the instrument correspond with teacher assessment of the child's performance. The individual doing the assessment may or may not agree with the behavioral standards of the peer group, the family, or the teacher. Such agreement is not necessary, so long as the measure accurately measures the adaptive fit of the child in the system, whatever its norms.

The reliability of the direct measures of performance in the social system i.e. teacher ratings, peer ratings, etc. is determined by inter-rater reliability when reports from two different informants from within the system are correlated.

The definition of test "bias" which is appropriate within this model is that propounded by Cleary et al. (1975). If the percent reduction in error, slope of the regression lines, and intercepts of the regression lines are similar for two racially or culturally different groups, then the measure is making racially and culturally nondiscriminatory predictions. If, however, any one of the three criteria are not met, predictions will be discriminatory. The direction of the discrimination will vary, depending on the specific relationships between the correlation coefficients, slopes of the regression lines, and intercepts of the regression lines in the particular case.

It is useful, in this context, to differentiate between direct and indirect measures of role performance. Direct measures of role performance would be evaluations of the individual's role performance by other members of the group. Indirect measures of role performance would be measures based on information from someone who is not a member of a group but an observer, measures based on self-reports and measures based on observations by a nonmember of the group. For example, sociometric ratings completed by members of the peer gang would be a direct measure of the extent to which a child was performing in a manner acceptable to the group. Information from a mother or a teacher on the child's adaptive fit to the peer gorup would be an indirect measrue. A teacher's rating of

124

a child's performance as a student or grades given a child by the teacher would be a direct measure of the child's role performance in the classroom. An academic acheivement test, an aptitude test, the parent's report of the child's performance in school would all be indirect measures of student role performance. In general, direct measures are used as the criteria against which an indirect measure is "validated." Academic achievement and aptitude tests are "validated" by correlating them with teacher ratings or teacher grades.

The General Intelligence Model. provides information on the child's learning potential which is useful primarily in making prognostic statements: Components C and E. Estimates of the child's "intelligence" can be made *if* the assumptions of the inferential model for making inferences about "potential" or "intelligence" have been met by taking into account the child's sociocultural background and the child's physical disabilities through using one of the approaches discussed in Section 2 of this paper. Prognostic statements in Component C, prognosis without intervention, will be more positive if the child's estimated learning potential is high rather than low. Likewise, educational objectives and time lines for Component E will be modified, up or down, depending on estimates of the child's learning potential.

In order to make inferences about "general intelligence" which are not racially or culturally discriminatory, it is necessary to take into account the child's sociocultural background and the extent to which the standard norms may or may not be appropriate for estimating the potential of a particular child. Information on sociocultural background can also be used, independently, as a source of Type A information. Knowledge of the child's family background can be used to inform the Assessment team of the gap between the sociocultural background of the child and the culture of the school. Such information will assist them in determining the magnitude of the problem faced by the child in attempting to bridge that gap. It can also provide insight into possible avenues for assisting the child through developing cooperative arrangements with the family.

The Assessment Team

In the guidelines and related documents for Public Law 94-142, two general proposals are made to assure that assessment data is not misused and misinterpreted: (1) training personnel to use the tests in the manner intended by the person who developed the test and (2) placing responsibility for educational decisions in a multidisciplinary planning conference which will include "persons knowledgeable about the child, the meaning of the evaluation data, and the placement options."

When all three assessment models are used to develop diagnostic constructs, as depicted in Figure 3, the planning conference or assessment team must include

persons from a variety of disciplines. An optimal configuration would include the following persons.

1. Educational psychologist or school psychologist. The person in this role would be trained in the administration of individual "intelligence" tests and would be responsible for making inferences about "general intelligence" within the general intelligence or pluralistic model. He or she could be involved in collecting and interpreting data within the Social Adaptivity Assessment model.

2. Special education personnel, resource teacher, educational diagnostician. The person in this role would be one who has been trained in administering and interpreting edumetric tests and would have specific knowledge and expertise in developing Individual Educational Plans.

3. School nurse. The medical model measures, both Type A and B information, would be provided by the school nurse: screening for visual or auditory impairments, sensorimotor skills, physical coordination, health history, and so forth. In addition, the school nurse would be responsible for medical interpretation of the health history and developmental history. He or she may or may not be involved in collecting the data for the health and developmental histories.

4. School social worker, counselor, or visiting teacher. The person in this role would be responsible for family contacts. Ordinarily this individual would collect the data from the family on the child's adaptive behavior in non-school settings, would collect data on the socio-cultural characteristics of the family, and so forth. They should also be trained to interpret adaptive behavior information from the family and to work with the family in implementing the Individual Educational Plan.

5. Classroom teacher. In addition to providing edumetric data on the child's performance in the classroom, the teacher would be directly involved in decisions concerning the elements of the Individual Educational Plan for the child and procedures for implementing that plan.

6. Specialists. In those cases requiring assessment in special areas, such as audiometry, speech, or vision, specialists in the appropriate areas would be added to the assessment team. A physician would be added when screening using medical model measures indicates the need for a more thorough medical review.

7. The child's parents and/or a community advocate. Persons from the child's family should participate by providing adaptive behavior information on the child's performance outside the school and the child's health and developmental history. They would also be involved in decisions relating to the Individual Educational Plan and its implementation.

8. Whenever feasible, the child should be included in the assessment team. To the maximum extent possible, the child should understand the nature of the assessment process, should be informed of the findings, and should be a party to the decisions made concerning his or her educational future.

CHAPTER 4: EVALUATION CRITERIA FOR PROTECTION IN ASSESSMENT

Criteria for evaluating the extent to which a particular jurisdiction is meeting the standards for protection in assessment described in this paper can be viewed at two levels: the level of the individual child being assessed and the level of the school district or other jurisdiction. If the level of the individual child is described, that data can be aggregated to produce an assessment at the institutional or district level. Each will be treated separately.

Public Law 94-142 suggests five major approaches to protection in evaluation: (1) multidimensional assessment, (2) determination of the validity of measures relative to the purposes for which they are being used, (3) using personnel trained to adminster and interpret measures in a manner congruent with the intentions of the person or persons who developed each measure, (4) determining whether the measure is racially or culturally discriminatory, and (5) the use of the multidisciplinary team.

Protection in Assessment Procedures: Level of Individual Child

Multidimensional Assessment.

The first stage in planning the assessment of a child is making a series of decisions concerning the scope or comprehensiveness of the assessment procedures required in a particular case. Ultimately, the range of measures secured must depend upon the professional judgement of the person in charge of planning the assessment. In some cases only partial diagnostic constructs will be needed to devise an appropriate intervention. In other cases, a complete diagnostic construct will be necessary. Following are some decision rules which can be used in planning the scope of a particular assessment.

(1) Is the evaluation a "preplacement" evaluation or a "re-evaluation"? A "re-evaluation" will ordinarily focus more specifically on the educational, behavioral, and medical objectives outlined in the original IEP and the choice of assessment instruments will be guided by those objectives. The range of measures in the original evaluation – called the "preplacement" evaluation in P.L. 94-142 – will be determined by two major factors: the nature of the "presenting problem(s)" as described by the teacher, parent, or other person who referred

127

the child and the restrictiveness of the "placement" or intervention contemplated for the child.

(2) Does the "presenting problem(s)" as described in the referral relate to medical model, social adaptivity model, or general intelligence model questions? The assessment battery should include evaluation in any or all of the three domains covered by the 3 models if they are mentioned as problem areas by persons knowledgeable about the child. For example, if the "presenting problem(s)" relate to academic difficulty, then edumetric measures based on the social system model are needed. If they relate to interpersonal relations with peers, then a measure of adaptive fit to the peer group based on the social system model is indicated. If the "problem" relates to vision and/or hearing and/or motor coordination etc. then medical model measures are necessary.

(3) The second parameter to be considered in deciding the scope of an assessment is the restrictiveness of the contemplated intervention. The general decision-rule is that the more restrictive the contemplated intervention the more complete the diagnostic construct must be which supports that intervention. Operationally, we can distinguish roughly 8 levels of restrictiveness of educational settings along a continuum.

Level 1: Regular Classroom Assignment: Special Intervention by the Classroom Teacher Only — No Removal from Regular Class.

Level 2: Regular Classroom Assignment: Special Intervention by Ancillary personnel such as tutors, resource teachers, specialists, etc. — No removal from regular classroom.

Level 3: Regular Classroom Assignment: Special Intervention by ancillary personnel involving removal from the regular classroom for 1 to 8 hours per week.

Level 4: Regular Classroom Assignment: Special Intervention by ancillary personnel or special education specialists involving removal from the regular classroom for 9 to 15 hours per week.

Level 5: Special Education Assignment: Education primarily in a special education setting with some activities integrated with the regular classroom, such as music, physical education, art, etc.

Level 6: Special Education Assignment: Completely self-contained educational program located on the regular "neighborhood" school campus.

Level 7: Special Education Assignment: Located in a separate school for special education.

Level 8: Placement in an Institutional Setting outside the community.

In general, interventions at Levels 1, 2, and 3 would not ordinarily require developing a complete diagnostic construct. Partial constructs involving Edumetric measures or "medical" type interventions (glasses, hearing aide) or short term minor interventions (as in the case of speech therapy for an articulation

128

problem) would suffice at these lower levels.

However, any intervention at Level 4 or higher would be sufficiently restrictive as to require the development of a complete diagnostic construct ABCDE. In completing the suggested forms for evaluating whether a child has had protection in evaluation procedures, NR is a code which stands for Not Relevant. It indicates only a Level 1, 2, or 3 intervention is contemplated *and* that particular assessment is not mentioned in the referral. For any evaluation in which the intervention goes beyond Level 3, *all* aspects of the Medical Model, Social Adaptivity Model, and General Intelligence Model should be measured, regardless of whether they are mentioned in the initial referral as problem areas.

Validity of Measures.
On the charts for reporting the quality of an individual assessment in meeting the criteria for protection in evaluation, each measure is rated for how well it has been "validated." Since the definition of "validity" varies with the assessment model, the rater must use the definition appropriate for each model. Four levels of rating are suggested: Well validated, adequately validated, poorly validated, and not valid or invalid.

Training of Personnel.
The best of measures can be invalidated if they are improperly administered or improperly interpreted. Thus, the rating of the qualifications of the person who administered, scored, and interpreted the findings for a particular measure is an important aspect of protection in assessment. Again four levels of rating are suggested: Well trained according to procedures suggested by the person or persons who developed the measure; Adequately trained; Poorly trained; and Untrained. The evaluation of training includes a judgement as to whether the person is interpreting scores on the measure in the manner intended by the person or persons who developed the measure. If such is not the case, the rating would be "Poorly trained", regardless of credentials held or the source of the training.

Evidence that the Measure is Racially and Culturally Non-Discriminatory.
Since each assessment model has a different definition of what is "racially and culturally non-discriminatory" and how that parameter is to be tested within each model, the person making the judgement on whether a particular measure is "non-discriminatory" will have to be familiar with the definitions and tests appropriate to each model. Again, four levels of judgement are suggested ranging from Well Documented, Adequately Documented, Poorly Documented, to No Documentation. The type of documentation needed will vary with each model.

Multidisciplinary Assessment Team.
The rating for the multidisciplinary team is based on the assumption that involvement at the level of both data collection and interpretation is the highest

129

level of involvement. For example, a social worker might be used to collect family interview data but not participate on the assessment team or be asked to interpret the interview data. This relationship would indicate a lower level of contribution than if the social worker not only collected family data but interpreted it and participated in helping to develop the IEP. Even the most minimal assessment would involve the teacher, parent, and child at some level. Therefore, the Not Relevant response is not applicable for these persons in rating how multidisciplinary the assessment team is.

Summary Score

Because some types of evaluation may not be relevant for all children, an average score on the relevant ratings is used as the summary score rather than a simple total. Thus, the summary score consists of the total number of points given each dimension measured summed to get an overall total which is divided by the number of dimensions rated. The minimal criteria would be an average rating of "2" or higher.

In addition, the ratings given individual dimensions should be examined. Any ratings of 0 or 1 on any aspect of any measure makes the measure suspect if a measure is poorly validated or not validated; is administered by poorly or untrained personnel; or there is little or no documentation that it is racially or culturally non-discriminatory, then that aspect of the assessment process should be repeated with different measures or better trained personnel, as the case may be.

REPORT OF MEDICAL MODEL MEASURES ADMINISTERED

Dimension Measured	Name of Test	Person Administering	Was Dim. Tested?	How Well Validated?	Trained Personnel?	Evidence Non-Discr?	Total (Enter NR's "Not Relevant")
1. Physical Dexterity			1 0 NR	3 2 1 0	3 2 1 0	3 2 1 0	
2. Visual-Motor Coor.			1 0 NR	3 2 1 0	3 2 1 0	3 2 1 0	
3. Visual Acuity			1 0 NR	3 2 1 0	3 2 1 0	3 2 1 0	
4. Auditory Acuity			1 0 NR	3 2 1 0	3 2 1 0	3 2 1 0	
5. Health History			1 0 NR	3 2 1 0	3 2 1 0	3 2 1 0	
6. Medical History			1 0 NR	3 2 1 0	3 2 1 0	3 2 1 0	
7. Other (specify)			1 0 NR	3 2 1 0	3 2 1 0	3 2 1 0	

Total (Range 0-70) Relative Total = Total / (7 - Total NR's) = Average Rating Per Measure Scored

Instructions for Completing Form:

Was Dimension Tested? 1 = Yes; 0 = No; NR = Not Relevant. Not Relevant should be circled only if the dimension is not mentioned as part of "presenting problem" and only a Level 1, 2, or 3 intervention in contemplated. Otherwise, circle "0" for Not Administered".

How Well Is Measure Validated? 3 = Well Validated i.e. many intercorrelations other Medical Model measures on different populations — not correlated sociocultural factors; 2 = Adequately Validated i.e. some studies on different populations; 1 = Poorly Validated i.e. few studies; 0 = Not validated i.e. no studies or negative findings.

How Well Trained was person administering test? 3 = Well trained; 2 = Adequately trained; 1 = Poorly trained; 0 = Not trained

What is Evidence the Measure is Racially and Culturally Non-Discriminatory? 3 = Well documented; 2 = Adequately documented; 1 = Poorly documented; 0 = No documentation. Evidence needed: Scores not correlated with sociocultural factors; scores are transcultural; measures measure organism status not learned behavior.

Criteria: All scores for individual ratings on four above categories should be 2 or 3 for all administered measures. Average Rating Per Measure Scored should be 2 or higher.

REPORT OF SOCIAL ADAPTIVITY OR SOCIAL SYSTEM MEASURES ADMINISTERED

PSYCHOMETRIC MEASURES

Dimension Measured	Name of Test	Person Administering	Was Dim. Tested?	How Well Validated?	Trained Personnel?	Evidence Non-Discrim.?	Total (Enter NR if not Rel.)
1. Academic Ach. Test			1 0 NR	3 2 1 0	3 2 1 0	3 2 1 0	
2. Nonacad. Sch. Roles			1 0 NR	3 2 1 0	3 2 1 0	3 2 1 0	
3. Family Roles			1 0 NR	3 2 1 0	3 2 1 0	3 2 1 0	
4. Peer Group Roles			1 0 NR	3 2 1 0	3 2 1 0	3 2 1 0	
5. Community Roles			1 0 NR	3 2 1 0	3 2 1 0	3 2 1 0	
6. Economic Roles			1 0 NR	3 2 1 0	3 2 1 0	3 2 1 0	
7. Self-Help/Mainten.			1 0 NR	3 2 1 0	3 2 1 0	3 2 1 0	
8. Student Role (Teach.)			1 0 NR	3 2 1 0	3 2 1 0	3 2 1 0	
9. Other (Specify)			1 0 NR	3 2 1 0	3 2 1 0	3 2 1 0	
Total (Range 0-90). Average Rating Per Measure Scored = Total ÷ (9 - Total NR's) =							

Instructions for Completing Form:

Was Dimension Tested? 1 = Yes 0 = No; NR = Not Relevant. Not relevant should be circled only if the dimension is not mentioned as part of the presenting problem and only a Level 1, 2, or 3 intervention is contemplated.

How Well Is Measure Validated? 3 = Well Validated (i.e. meets all criteria on several ethnic groups; 2 = Adequately Validated (i.e. meets most criteria on several ethnic groups; 1 = Poorly Validated (i.e. meets few criteria or only white students studied); 0 = Not Validated (i.e. no studies or negative findings.

How Well Trained was Person Administering Test? 3 = Well trained by producer of test; 2 = Adequately trained; 1 = Poorly trained; 0 = Not Trained

Evidence Measure is Racially and Culturally Non-Discriminatory? 3 = Clear evidence predictions unbiased; 2 = Some evidence predictions unbiased; 1 = Little evidence predictions unbiased; 0 = No evidence. Direct Measures scored 3.

Criteria: All scores for ratings of individual tests should be 2 or higher if test was administered. When scored NR, there must be clear evidence that the test is, indeed, not relevant. Average Rating Per Measure scored should be 2 or higher.

REPORT OF SOCIAL ADAPTIVITY OR SOCIAL SYSTEM MEASURES ADMINISTERED

Ethometric Measures

Dimension Measured	Name of Test	Person Administering	Was Dim. Tested?	How Well Validated?	Trained Personnel?	Evidence Non-Discrim.?	Total (Enter NR if Not Rel.)
1. "Diagnostic" Tests			1 0 NR	3 2 1 0	3 2 1 0	3 2 1 0	
2. Teacher Con. Tests			1 0 NR	3 2 1 0	3 2 1 0	3 2 1 0	
3. Teacher Report/Obsr.			1 0 NR	3 2 1 0	3 2 1 0	3 2 1 0	
4. Teacher grades GPA			1 0 NR	3 2 1 0	3 2 1 0	3 2 1 0	
5. Other (Specify)			1 0 NR	3 2 1 0	3 2 1 0	3 2 1 0	
Total (Range 0-50) Average Rating Per Measure Scored = Total / (5 - NR's) =							

Instructions for Completing Form:

Was Dimension Tested? 1 = Yes; 0 = No; NR = Not Relevant. An XX is inserted for Dimensions 2, 3, and 4 on chart because ethometric data from the teacher is always relevant, even in a minimal assessment.

Scoring of the other categories is the same as described under PSYCHOMETRIC MEASURES.

REPORT OF GENERAL INTELLIGENCE OR PLURALISTIC MODEL MEASURES ADMINISTERED

Dimension Measured	Name of Test	Person Administering	Was Dim. Tested?	How Well Validated?	Trained Personnel?	Evidence Non-Discrim.?	Total (Enter NR if not Rel.)
1. Global "Intelligence"			1 0 NR	3 2 1 0	3 2 1 0	3 2 1 0	
2. Verbal "Intelligence"			1 0 NR	3 2 1 0	3 2 1 0	3 2 1 0	
3. Performance "Intell."			1 0 NR	3 2 1 0	3 2 1 0	3 2 1 0	
4. Other (Specify)			1 0 NR	3 2 1 0	3 2 1 0	3 2 1 0	
Total (0-40) Average Rating Per Measure Score = Total / (4 - NR's) =							

Instructions for Completing Form:

Was Dimension Tested? Same instructions as on earlier forms.

How Well is Measure Validated? 3 = Clearly meets all assumptions of inferential model for inferring "intelligence" 2 = Partially meets assumptions of inferential model 1 = Questionable if meets assumptions. 0 = No evidence it meets assumptions or evidence that it does *not* meet assumptions.

How Well Trained was Person Administering Test? Same instructions as on earlier forms.

Evidence Measure is Racially and Culturally Non-Discriminatory? 3 = Clear evidence on several ethnic groups 2 = Some evidence on a few ethnic groups 1 = Questionable evidence - mainly speculation 0 = No evidence it is non-discriminatory in making inferences or findings are that there is discrimination.

REPORT ON ASSESSMENT TEAM UTILIZED IN ASSESSMENT

Team Member	*Collect and Interpret*	*Collect Data Only*	*Not Involved*	*Not Relevant*
1. Educational Psychologist or similarly trained person	2	1	0	NR
2. Nurse or other medically trained person	2	1	0	NR
3. Social worker, visiting teacher, of similar person	2	1	0	NR
4. Classroom teacher	2	1	0	XX
5. Special Education Teacher, resource teacher, similarly trained person	2	1	0	NR
6. Parent(s) or parent advocate	2	1	0	XX
7. The child being assessed	2	1	0	XX

Total (Range 0-14) _____

Instructions for Completing Form:

Circle the number that best describes the degree of involvement of the type of person described in each of the seven categories. In cases in which the referral was for a very specific need (vision test, hearing test, etc.) the involvement of certain persons might not be "relevant" to the problem and NR would be circled. However, in *all* cases the teacher, parent, and child would be involved at some level. Hence, NR is not an appropriate response for those individuals.

2 = Involvement both at the level of collecting and interpreting the data i.e. full involvement on the assessment team.

1 = Involvement at the level of collecting data – either by providing the data (as in the case of the parent, teacher, or child) or in collecting it (as in the case of the social worker conducting the parent interview, the nurse administering the physical dexterity tasks). The category indicates that the individual does not get involved in the assessment planning conference in which data are interpreted, the diagnostic construct built, and the IEP developed.

0 = No involvement at any level.

NR = Not relevant. This category should not be used if the presenting problem was in the area of that particular person's professional expertise *or* if the contemplated placement is at a higher level of restrictiveness than level 3.

135

FORM FOR SUMMARIZING DISTRICT LEVEL DATA ON PROTECTION–IN–EVALUATION PROCEDURES

It is suggested that a district report would consist of the percent of the cases evaluated in the district during a given period which have met the minimal criteria on each of dimensions within each model. The summary form might be similar to the following format.

Multiple Measures

Ratings on this characteristic would be divided into two types of cases those for which partial diagnostic constructs were developed and those with full diagnostic constructs.

	100%	*75%*	*50%*	*25%*	*0%*
1. What percent of the cases for which partial constructs are indicated were tested for specific problems indicated in initial referral?	4	3	2	1	0
2. What percent of the cases for which the restrictiveness of the intervention was Level 4 or higher were administered all the measures needed for a full diagnostic construct?	4	3	2	1	0

Validity of Measures

1. What percent of the cases for which full constructs were required had an average rating of 2 or higher on validity of medical model measures used?	4	3	2	1	0
2. What percent had an average rating of 2 or higher on social system measures used?	4	3	2	1	0
3. What percent had an average rating of 2 or higher on general intelligence model measures used?	4	3	2	1	0

Trained Personnel

1. What percent of the cases had an average rating of 2 or higher on the training of personnel administering medical model measures?	4	3	2	1	0
2. What percent had an average rating of 2 or higher on the training of personnel administering measures in the social adaptivity model?	4	3	2	1	0
3. What percent had a rating of 2 or higher on training of personnel administering general intelligence model measures?	4	3	2	1	0

Evidence Tests are Non-Discriminatory	*100%*	*75%*	*50%*	*25%*	*0%*
1. What percent of the cases which used medical model measures had an average rating of 2 or higher when judged by the criteria appropriate for "racially and culturally non-discriminatory" testing within a medical model?	4	3	2	1	0
2. What percent of the cases which used social adaptivity measures had an average rating of 2 or higher when judged by the criteria appropriate for "racially and culturally non-discriminatory" testing within the social adaptivity model?	4	3	2	1	0
3. What percent of the cases which made inferences about general intelligence and or learning potential had an average rating of 2 or higher when judged by the criteria for making "racially and culturally non-discriminatory" inferences within the general intelligence model?	4	3	2	1	0
Multidisciplinary Team					
1. What percent of the cases had all 7 types of professionals or participants listed as members of the planning conference involved at the data collection level (1) or higher?	4	3	2	1	0

137

Summary

The forms presented in this paper should be regarded as suggestive rather than final. Collecting data of this sort is a complex task and considerable thought, pre-testing, and revision would be required before any set of forms could be utilized. It is difficult to establish a minimal criteria that is less than the optimal assessment because in the case of any given child the one lapse in an appropriate evaluation may be the lapse which is crucial in the mis-labeling or mis-educating of that child. Therefore, each jurisdiction should be urged to fulfill all aspects of the procedures required to protect the child from the effects of an inappropriate or discriminatory evaluation.

REFERENCES

Bernal, E. M., Jr. A response to "educational uses of tests with disadvantaged subjects." *American Psychologist*, January 1975, pp. 93-95.

Bersoff, D. N. Silk purses into sows' ears: The decline of psychological testing and a suggestion for its redemption. *American Psychologist*, 1973, *28*, 892-899.

Breland, H. M., et al. *The cross-cultural stability of mental test items: An investigation of response patterns for ten sociocultural groups.* Princeton, New Jersey: Educational Testing Service, 1974.

Budoff, M. *Measuring learning potential: An alternative to the traditional psychological examination.* Paper presented at the First Annual Study Conference in School Psychology, Temple University, Philadelphia, June 1972.

Carver, Ronald P. Two dimensions of tests: Psychometric and edumetric. *American Psychologist*, July 1974, pp. 512-518.

Cattell, H. *Culture fair tests for measuring intelligence.* Institute for Personality and Ability Testing, 1973.

Cleary, T. Anne; Humphreys, Lloyd G.; Kendrick, S. A.; and Wesman, Alexander. Educational uses of tests with disadvantaged students. *American Psychologist*, January 1975, pp. 15-40.

Coyle, F. A. Another alternate wording on the WISC. *Psychological Reports*, 1965, *16*(3), pt. 2, 1276.

Cromwell, Rue L.; Blashfield, Roger K.; and Strauss, John S. Criteria for classification systems. In Nicholas Hobbs (Ed.), *Issues in the classification of children* (Vol. 1). San Francisco: Jossey-Bass, 1975.

Cronbach, L. J. Five decades of public controversy over mental testing. *American Psychologist*, 1975, *30*, 1-14.

DeAvila, E. A., and Havassy, B. E. Piagetian alternative to IQ: Mexican-American study. In Nicholas Hobbs (Ed.), *Issues in the classification of children* (Vol. 2). San Francisco: Jossey-Bass, 1975.

Dennis, W. Goodenough scores, art experience, and modernization. *Journal of Social Psychology*, April 1966, *68*, 211-228.

Doll, E. A. *Measurement of social competence: A manual for the Vineland social maturity scale*. Circle Pines, Minn.: American Guidance Service, 1953.

Eells, K., et al. *Intelligence and cultural differences*. Chicago: University of Chicago Press, 1951.

Garber, Howard L. Intervention in infancy: A developmental approach. In Michael J. Begab and Stephen A. Richardson (Eds.), *The mentally retarded and society: A social science perspective*. Baltimore: University Park Press, 1975.

Grossman, Herbert J. (Ed.). *Manual on terminology and classification in mental retardation* (Special Publication Series No. 2.) Washington, D. C.: American Association on Mental Deficiency, 1973.

Harris, D. B. *Children's drawing as measures of intellectual maturity*. New York: Harcourt Brace Jovanovich, 1963.

Irvine, S. H. Toward a rationale for testing attainments and abilities in Africa. *British Journal of Educational Psychology*, February 1966, *36*: 24-32.

Jackson, G. D. On the report of the ad hoc committee on educational uses of tests with disadvantaged students: Another psychological view from the Association of Black Psychologists. *American Psychologist*, January 1975, pp. 88-93.

Jencks, Christopher; Smith, Marshall; Acland, Henry; Bane, Mary Jo; Cohen, David; Gintis, Herbert; Heyns, Barbara; and Michelson, Stephan. *Inequality: A reassessment of the effect of family and schooling in America*. New York: Basic Books, 1972.

MacKay, R. Standardized tests: Objective/objectified measures of "competence." In A. V. Cicourel (Ed.). *Language and school performance*. New York: Academic Press, 1974.

Mehan, H. Assessing children's language-using abilities: Methodological and cross-cultural implication. In M. Armer and A. D. Brimshaw (Eds.), *Comparative social research: Methodological problems and strategies*. New York: John Wiley, 1973.

Mercer, Jane R. Who is normal: Two perspectives on mild mental retardation. In E. G. Jaco (Ed.), *Patients, physicians, and illness* (Rev. Ed.). Glencoe, Ill.: Free Press, 1972.

140

Mercer, Jane R. *Labeling the mentally retarded*. Berkeley and Los Angeles, University of California Press, 1973.

Mercer, Jane R. Sociocultural factors in educational labeling. In Michael J. Begab and Stephen A. Richardson (Eds.), *The mentally retarded and society: A social science perspective*. Baltimore: University Park Press, 1975.

Mercer, Jane R. *Theoretical constructs of adaptive behavior: Movement from a medical to a social-ecological perspective* (Technical Report No. 4). Austin: Texas Regional Resource Center, 1977.

Mercer, Jane R., and Brown, Wayne C. Racial differences in IQ: Fact or artifact? In C. Senna (Ed.), *The fallacy of IQ*. New York: Third Press, 1973.

Mercer, Jane R., and Lewis, June F. *System of Multicultural Pluralistic Assessment (SOMPA)*. New York: The Psychological Corporation, 1978.

Moran, R. E. Observations and recommendations on the Puerto Rican version of the Wechsler Intelligence Scale for Children. *Pedagogia*, Rio Piedros, 1962, *10*, 89-98.

Raven, J. C. *Guide to the standard progressive matrices*. London: H. K. Lewis, 1960.

Roth, D. R. Intelligence testing as a social activity. In A. V. Cicourel (Ed.), *Language use and school performance*. New York: Academic Press, 1974.

Scheff, Thomas J. *Being mentally ill: A sociological theory*. Chicago: Aldine, 1966.

Schwarz, P. A. Adapting tests to the cultural setting. *Educational and Psychological Measurement*, 1963, *23* (4).

Smith, M. W. Alfred Binet's remarkable questions: A cross-national and cross-temporal analysis of the cultural biases built into the Stanford-Binet Intelligence Scale and other Binet tests. *Genetic Psychology Monographs*, May 1974.

Terman, Lewis M., and Merrill, Maud A. *Stanford-Binet Intelligence Scale: Manual for the third revision form L-M*. Boston: Houghton Mifflin, 1960.

Webster's Third New International Dictionary. Springfield, Mass.: G. & C. Merriam, 1966.

Wechsler, David. *WISC-R: Manual for Wechsler Intelligence Scale for Children, Revised*. New York: The Psychological Corporation, 1974.

Wesman, Alexander G. Intelligent testing. *American Psychologist*, April 1968, *23* (4), pp. 267-275.

Williams, Robert L. The BITCH-100: A culture specific test. *The Journal of Afro-American Issues*, 1975, *3* (1), 103-116.

SECTION III

Implementing the Protection in Evaluation Procedures Provision of P.L. 94-142

James E. Ysseldyke

YSSELDYKE, JAMES E. Dr. Ysseldyke is Associate Professor of School Psychology and Director of the Institute for Research on Learning Disabilities at the University of Minnesota. A recipient of the Doctorate in School Psychology from the University of Illinois, he has taught special education classes and been a school psychologist. Professor Ysseldyke is co-author, with John Salvia of a text entitled Assessment in Special and Remedial Education, and has published numerous journal articles and book chapters on assessment. In 1973 he received the Lightner Witmer Award from the Division of School Psychology of the American Psychological Association for his research on diagnostic-prescriptive teaching.

144

INTRODUCTION

The recent and significant revisions in public policy on the education of handicapped children are reflected in the provisions of Public Law 94-142, the Education for All Handicapped Children Act. The law is designed to meet four major purposes, described by Ballard and Zettel (1977), as follows:

1. to guarantee that special educational services are available to children who need them,
2. to assure that decision-making regarding provision of services to handicapped students is both fair and appropriate,
3. to establish clear management and auditing requirements and procedures for special education at all levels of government, and
4. to provide federal funds to assist states in educating handicapped children and youth.

Public Law 94-142, along with Public Law 93-112, the Vocational Rehabilitation Act of 1973, represent significant entrance of legislation into the special education arena. This paper focuses on one special provision in Public Law 94-142, the "Protection in Evaluation Procedures" provision. This provision of the Law (Section 615-5c) specifies that states and their localities will develop:

> Procedures to assure that testing and evaluation materials and procedures utilized for the purposes of evaluation and placement of handicapped children will be selected and administered so as not to be racially or culturally discriminatory. Such materials or procedures shall be provided and administered in the child's native language or mode of communication, unless it clearly is not feasible to do so, and no single procedure shall be the sole criterion for determining an appropriate educational program for a child.

Local and State education agencies are required to demonstrate compliance with these "Protection in Evaluation Procedures" provisions of the law. Specific rules and regulations for implementation of the PEP provisions were published in the *Federal Register*, August 23, 1977 (pp. 42474-42518). These rules and regulations specify that:

---

1. Before any action is taken with respect to the initial placement of a handicapped child in a special education program, a full and individual evaluation of the child's educational needs must be conducted in accordance with the requirements of rule 121a.532 (Rule 121a.531).
2. State and local education agencies shall insure, at a minimum, that:
   (a) Tests and other evaluation materials:
      (1) Are provided and administered in the child's native language or other mode of communication, unless it is clearly not feasible to do so;
      (2) Have been validated for the specific purpose for which they are

used; and

(3) Are administered by trained personnel in conformance with the instructions provided by their producer;

(b) Tests and other evaluation materials include those intended to assess specific areas of educational need and not merely those which are designed to provide a single general intelligence quotient;

(c) Tests are selected and administered so as best to ensure that when a test is administered to a child with impaired sensory, manual, or speaking skills, the test results accurately reflect the child's aptitude or achievement level or whatever other factors the test purports to measure, rather than reflecting the child's impaired sensory, manual, or speaking skills (except where those skills are the factors which the test purports to measure);

(d) No single procedure is used as the sole criterion for determining an appropriate educational program for a child; and

(e) The evaluation is made by a multi-disciplinary team or group of persons, including at least one teacher or other specialist with knowledge in the area of suspected disability.

(f) The child is assessed in all areas related to the suspected disability, including where appropriate, health, vision, hearing, social and emotional status, general intelligence, academic performance, communicative status, and motor abilities (20 U.S.C. 1415 (b) (2) (B) (121a.532 a-f)).

This paper addresses the general issue of assuring that SEAs and LEAs have a means of evaluating the extent to which their assessment procedures are in compliance with the "PEP" provision and its accompanying rules and regulations. Part I of the paper is a brief overview of the ways in which the concepts of nondiscriminatory assessment and bias in assessment have been addressed in the professional literature. Part III is an overview of factors which must be considered if we are to address satisfactorily the issue of bias in assessment. Part II ends with this author's interpretation of nondiscriminatory assessment and a statement of the rationale for the position developed.

Part III is an outline of factors considered in developing specific criteria for evaluating effective implementation of the PEP provision, while Part IV is a set of criteria or standards for use in evaluating LEA implementation of the PEP provision.

TREATMENT OF THE CONCEPTS OF NONDISCRIMINATORY ASSESSMENT AND BIAS IN ASSESSMENT: A LOOK AT THE PROFESSIONAL LITERATURE

The issue of bias in assessment has been with us for a long time and has been dealt with variously in the professional literature. Early research on bias in

assessment began with the observation that when norm-referenced tests were administered to students from majority and minority groups, the members of the minority group *on the average* earned lower scores. The fact that, in general, average scores earned by *groups* of minority students tend to be lower than those earned by groups of nonminority students led to numerous speculations regarding the reasons for the observed differences. While some investigators have argued that observed differences between groups for the most part reflect genetic differences between groups, others have argued that observed differences are primarily due to differential environmental effects. The "positions of choice" these days appears to be an "interactionist position" in which the performance of an individual on a test is viewed as a function of an interaction between genetic and environmental influences. The nature-nurture debate has produced, within the past decade, a plethora of theorizing and numerous empirical investigations. Investigators have debated the concept of "intelligence" (Cattell, 1963, 1971; Elkind, 1969; 1969; Guilford, 1967; Humphreys, 1971; Merrifield, 1971; Vernon, 1969; Wechsler, 1971); and the relative contributions to intelligence of genetic and environmental variables (Bayley, 1965; Bereiter, 1969; Bijou, 1971; Bloom, 1964; Bodmer & Cavalli-Sforza, 1970; Burt, 1967; Butcher, 1968; Cattell, 1953, 1971; Cronbach, 1969; Crow, 1969; Dreger & Miller, 1968; Eckland, 1971; Eells, Davis, Havighurst, Herrick & Cronbach, 1951; Elkind, 1969; Erlenmeyer-Kimling & Jarvik, 1963; Ginsberg & Laughlin, 1971; Gordon, 1971; Greenfield, 1971; Hirsch, 1971; Hunt, 1961, 1969; Hunt & Kirk, 1971; Jensen, 1967, 1968a, 1968b, 1969a, 1969b, 1971; Kagan, 1969; Li, 1971; Vandenberg, 1971). A large number of studies have been designed to investigate the fairness of tests by comparing the performances of groups of students on norm-referenced tests (Boone & Adesso, 1974; Breland, et al., 1975; Butler, Coursey & Gatz, 1976; Gilmore, et. al., 1975; Goldman & Hewett, 1976; Guilliams, 1975; Hartlage & Lucas, 1976; Hennessey & Merrifield, 1976; Hoepfner & Strickland, 1972; Jensen, 1974, 1976; Kallignal, 1971; Matuszek & Oakland, 1972; McNeil, 1975; Mercer, 1972; Neal, 1975; Peck, 1973; Pfeiffer & Sedlacek, 1971; Ratusnik & Koenigsknecht, 1975; Reschly, et al., 1976; Rincon, 1976; Temp, 1971).

Investigations of group differences in performance on psychometric devices led other investigators to examine the fairness of specific items as used with members of minority groups (Angoff & Ford, 1971; Breland, 1974; Breland, et al., 1974; Durovic, 1975; Fishbein, 1975; Green, 1971; Green & Roudabush, 1976; Lord, 1976; Merz, 1976; Newland, 1973; Pine & Weiss, 1976; Rudner, 1977; Scheuneman, 1976; Smith, 1974; Tinsley & Dawes, 1972). Specific sub-components of the research on the extent to which specific tests and test items are biased against members of minority groups have been observed in research on linguistic bias (Bartel, Grill & Bryen, 1973; Berry & Lopez, 1977; Bryen, 1974; Johnson, 1973; Lefley, 1975; Matluck & Mace, 1973; Matluck & Mace-Matluck, 1975; Nathanson, 1975; Vasquez, 1972) and on sex bias (Diamond, 1976; Dwyer, 1976; Evans & Sperekas, 1975; Faggen-Steckler, et. al.,

147

1974; Harmon, 1973; Holland, 1976; Lockheed-Katz, 1974; Prediger & Hanson, 1976; Strassberg-Rosenberg & Donlon, 1975; Tittle, 1973, 1974; Tolor, 1975).

Research demonstrating differences between groups in performance on tests or test items, along with contentions that bias existed in selection and employment of people, led several psychologists to develop models of evaluating test fairness in their efforts to define the concept of culture fairness. Cole (1973) and Petersen and Novick (1976) have provided a useful conceptualization of six different models of fairness. These are summarized briefly as follows:

The Quota Model.

Research on differences in the performance of groups of students on tests has led to the viewpoint that a test is biased if it fails to identify proportions of individuals comparable to proportions in the general population. For example, if a test were used to select students for admission to a University, the test would be said to be biased if it resulted in the selection of a lower ratio of Blacks to whites than their porportions in the general population. If 11% of the general population are BLack; and the test resulted in the selection of a population in which only 8% of the students were Black, the test would be characterized as biased.

The Regression Model.

The regression model was originially proposed by Cleary (1968) who defined a biased test as follows:

> A test is biased for members of a sub-group of the population if, in the prediction of a criterion for which the test is designed, consistent nonzero errors of prediction are made for members of the subgroup. In other words, the test is biased if the criterion score predicted from the common regression line is consistently too high or too low for members of the subgroup. With this definition of bias, there may be a connotation of "unfair," partiuclarly if the use of the test produces a prediction that is too low (p. 115).

The Subjective Regression Model.

This model, also referred to as the Culture-Modified Criterion Model (Petersen & Novick, 1976) has been proposed by both Darlington (1971) and Linn (1972). Proponents of this model believe that fairness can be achieved only by combination of the Quota Model and the Regression Model. Accordingly, one first decides if there is some merit or value in selection of members of some

cultural group, or if there is potential harm in exclusion of the members of some group, and then sets his/her regression equations to account for desired representation in selection.

The Equal Risk Model.

The Equal Risk Model was proposed for use in industry by Guion (1966), and later by Einhorn and Bass (1971). According to Guion (1966), "Unfair discrimination exists when persons with equal probabilities of success on the job have unequal probabilities of being hired for the job." When assessment devices are used to select employees, and when persons with equal probabilities of success have unequal probabilities of selection due to the use of the devices, the devices are said to be unfair or biased. A test is fair, from this perspective, if it results in the selection of persons for employment or participation in an activitiy in a porportion equal to the proportion of those who succeed in that employment or activity. Applied to educational settings, a test would be said to be fair if it simply selected for entrance into college, for example, the same proportion of individuals from minority groups as could be expected to complete college successfully.

The Constant Ratio Model.

This model, proposed by Thorndike (1971), states that a test is fair when it results in the selection of the proportions of different cultural groups as would be achieved if the person doing the selection had available each subject's exact score on the criterion measure. According to Throndike (1971), in a fair selection procedure:

> The qualifying score on a test should be set at levels that will qualify applicants in the two groups in proportion to the fraction of the two groups reaching a specified level of criterion performance (p. 63).

The critical variable in this model is success on a criterion measure. Tests that fail to predict success for minority or majority persons are said to be biased. This model is called the constant ratio model because fairness is evidenced when the proportion of those selected to those successful is the same for any two groups.

The Conditional Probability Model.

This model, proposed by Cole (1973), again addresses the relationship between success on a test and success on a criterion measure. Cole states:

> For both minority and majority groups whose members can achieve a satisfactory criterion score (Y-Yo) there should be the same probability of selection regardless of group membership (p. 240).

Several investigators have reviewed the models of test fairness (Frazer, 1975; Hunter & Schmidt, 1976; Linn & Werts, 1971; McNemar, 1975; Petersen & Novick, 1974, 1976), and have concluded that there is little agreement among the several models. It is readily apparent that major measurement experts have been essentially *unable* to agree on a definition of a fair test, let alone identify a test that is fair for members of different groups. There is little agreement on the *concept* of nondiscriminatory assessment. As Petersen and Novick (1976) note:

> The Regression, the Constant Ratio, the Conditional Probability, the Equal Probability, the Equal Risk and the Culture-Modified Criterion Models are each explications of general concepts of what constitutes the fair use of tests in a selection situation. There seems to be nothing in the literature that clearly indicates when, if ever, one of the models is preferable to the other five models. Thus, the practitioner has no clear guidance in the choice of a culture-fair selection model. Further, we have suggested that the Constant Ratio, the Conditional Probability, the Equal Probability Models and their converses are *internally contradictory* (p. 23-24).

In addition to analyses of the fairness of specific tests and items we have witnessed a number of other concerns in the professional literature. There have been several analyses of litigation and legal issues (Linn, 1972; Nolte, 1975; Sharf, 1977; Weckstein, 1973), calls for culture specific assessment (Long & Anthony, 1974; Moran, 1974; Mukherjee, et al., 1976; Simon & Joiner, 1974; Williams, 1975), and discussions of the social and legal consequences of using tests to classify individuals (Epps, 1973; Hunt, 1972; Green, 1973; Kamin, 1973, 1975).

Cronbach (1975, 1976) has provided an excellent analysis of the socio-political nature of the arguments regarding testing. The considerable controversy regarding this topic has resulted in the recent publication of numerous position papers (Barnes, 1973; Bersoff & Ysseldyke, 1977; Cervantes, 1974; DeGeorge, 1975; Fitzgibbon, 1970; Flaugher, 1974; Franklin, 1974; Green, 1971; Humphreys, 1973; Jensen, 1974; McClelland, 1973; McNemar, 1975; Meeker & Meeker, 1973; Mercer, 1972; Messick & Anderson, 1970; National Association for the Advancement of Colored People, 1976; Northeast Regional Resource Center, 1976; Ratteray, 1974; Samuda, 1973; Scales & Smith, 1974; Scarr, 1977; Southwest Regional Resource Center, 1977; Weber, 1974; Williams, 1970, 1974; Zirkel, 1972).

Clearly, this nation and its researchers have vested considerable effort, time, and financial resources in attempts to develop or identify assessment devices that are not biased against members of racial or cultural groups.

150

THE RATIONALE FOR THIS PAPER

Review of Congressional testimony relevant to the "Protection in Evaluation Procedures" indicates an obviously much broader concern than simply with the fairness of tests and test items as used with members of minority groups. This broader concern is with abuse in the entire process of using assessment data to make decisions about pupils. It is this broader concern, abuse in decision-making, that this position paper addressed. Abuse is evident in many arenas relevant to assessment of children and includes: 1) inappropriate and indiscriminate use of tests, 2) bias in the assessment of handicapped children, and in identifying as handicapped those who are not, 3) bias throughout the decision-making process, and 4) bias following assessment.

I cite here several quotes from Senate reports that illustrate my reasons for stating that the "real concern" is with *abuse* in assessment and decision-making.

> The Committee is deeply concerned about practices and procedures which result in classifying children as having handicapping conditions when, in fact, they do not have such conditions. . . . At least three major issues are of concern with respect to problems of identification and classification: (1) the misuse of appropriate identifcation and classification data within the educational process itself; (2) discriminatory treatment as the result of the identifcation of a handicapping condition; and (3) misuse of identification procedures or methods which results in erroneous classification of a child as having a handicapping condition (Senate, Report No. 94-168, Education for All Handicapped Children Act, June 2, 1975, p. 26-29).

> The Committee is alarmed about the abuses which occur in the testing and evaluation of children, and is concerned that expertise in the proper use of testing and evaluation procedures falls far short of the prolific use and development of testing and evaluation tools. The usefulness and mechanistic ease of testing should not become so paramount in the educational process that the negative effects of such testing are overlooked (Senate, Report No. 94-168, Education for All Handicapped Children Act, June 2, 1975, pp. 26-29).

There is considerable disparity between ways in which professional educators have, to date, been addressing the issue of bias in assessment and the ways in which they can best address the issue. Efforts to engage in "fair" testing have been characterized by the following kinds of proposals (Mercer, 1977):

1. *Development of Culture Free Tests.* Many have proposed we engage in extensive efforts to develop Culture Free tests. Such efforts, historically, have been impossible because there is no culture free learning. Learning occurs in environmental contexts and, in fact, consists primarily of the inculcation of the culture.
2. *Development of Culture Fair Tests.* Efforts have been made to construct tests in which items are "balanced" so they represent multiple languages and

151

cultures. Such efforts have been unsuccessful. "Culture fair" tests have not demonstrated good predictive validity; they are poor predictors of success in a mono-cultural school system, and have been for the most part rejected by the dominant cultural group.

3. *Development of Culture-Specific Tests.* Efforts have been made, as noted earlier, to develop culture-specific tests (specific to the Black culture, Hispanic culture, etc). These tests, like culture fair tests, have demonstrated low predictive validity, and have been rejected by the dominant cultural group. Production of such devices has been difficult, because no one test will satisfactorily assess the heterogeneous group of children in any one cultural or ethnic group.
4. *Use of Piagetian Tests.* On many occasions those who seek fair assessment of children have proposed the use of Piagetian developmental scales. Such efforts have not resulted in fair assessment; specific items are as culturally dependent as are items on more traditional scales, and predictive validity is low.
5. *Linguistic Translation of Existing Tests.* Efforts to administer tests in children's native language have often consisted of translating existing tests into other languages. Translation changes item difficulty and destroys the applicability of existing norms (which unfortunately are too often used despite item translation). Once again, predictive validity for success in a monocultural school system has been low.
6. *Alteration of Administration Procedures.* When handicapped youngsters are assessed, assessors often try to achieve fairness by changing administration procedures. We witness, for example, the administration of verbal tests in sign language to deaf and hearing-impaired children. Nonstandardized administration procedures disallow the use of existing norms, and unless special population norms are constructed, norm-referenced interpretations are impossible.
7. *Training Children to Take Tests.* Some researchers and practitioners have advocated that children be trained to take tests prior to being assessed. Specific procedures have ranged from training in test-wiseness to task familiarization training. This is a viable way to eliminate or reduce observed score variance due to lack of familiarity with what is being required. The procedure is a time-consuming, but worthwhile alternative to traditional assessment procedures. It will not, however, reduce many aspects of bias in decision-making.
8. *Use of Pluralistic Norms for Existing Tests.* Mercer and Lewis (1978) developed a System of Multicultural Pluralistic Assessment (SOMPA). The system uses existing tests, but pluralistic norms. Separate regression equations are used to compute the estimated learning potential of children from Black, Anglo, and Chicano groups. One difficulty with such a procedure is in accounting for the extremely heterogeneous nature of any one cultural or ethnic group.

152

Professional decision-makers have repeatedly strived to identify fair assessment practices. Today, though, we very often observe SEA and LEA personnel engaged in efforts designed to identify *the* fair test for use with racial, cultural, or ethnic minority groups. Such efforts could go on for a very long time without producing progress toward non-discriminatory assessment. This fact was illustrated by Salvia and Ysseldyke (1978) in their overview of intellectual assessment. A description of that reasoning follows.

Intelligence tests, like any tests, are merely samples of behavior. Any student's performance on an intelligence test is a function of an interaction between the kind(s) of behavior(s) sampled by the test and the kinds of background experiences and opportunities that children have had in both formal and informal educational environments. Given the tremendous variation in background experiences with which children enter testing settings and the large number of different kinds of behaviors sampled by tests, the number of possible interactions is larger than we can even begin to realize (conservatively estimated by Salvia and Ysseldyke as greater than 1.35×10^{32} possible interactions). Educators can, and probably will, argue for a very long time about which of these interactions are "fair". We will make considerably more progress, I believe, by addressing something we *can* effect: bias in the entire process of decision-making. In establishing the position taken in this paper, it is necessary first to describe assessment and decision-making and then to describe the factors that I believe must be considered.

Assessment Defined

I view assessment broadly as the process of collecting data for the purpose of helping a professional make decisions about inidividuals. Assessment is not synonymous with testing; testing is simply one part of assessment. Assessment may include direct observation of inidividuals in natural environments, it may include the obtaining of data from others by means of interviews, and it may include the obtaining of both historical and current information by searching of records. Clearly, many different kinds of data are collected in the process of decision making; in its broadest sense, this data collection process is assessment.

Salvia and Ysseldyke (1978) identified five different kinds of decisions made in educational settings. Assessment plays a key role in the making of these decisions; data gathered by means of assessment are used in decision-making. The five kinds of decisions described by Salvia and Ysseldyke (1978) are briefly as follows.

1. *Screening.* In screening, data are collected for the purpose of helping professionals identify the extent to which a student's behavior differs from "normal" or "average" behavior. Students whose behavior is sufficiently

153

different from "normal" are typically referred for or identified as candidates for further assessment.

2. *Placement/Classification Decisions.* Assessment data are routinely collected in educational settings for the purpose of helping professionals decide how to classify students, for the purpose of declaring children eligible for special educational services, and as an aid in making placement decisions. Most state education agencies require that before children are placed in classes for the handicapped, they receive an individualized psychoeducational evaluation. Rules and regulations for P.L. 94-142 (sec. 121a.531) require that:

> Before any action is taken with respect to the initial placement of a handicapped child in a special education program, a full and inidividual evaluation of the child's educational needs must be conducted in accordance with the requirements of rule 121a.532.

Many different kinds of data are collected during individual evaluation, and these data are used to make placement, eligibility, or classification decisions. In the educational process, the appropriate identification of handicapping conditions must take place in order to assure that a child receives appropriate services designed to meet his or her needs. Identification must also take place to enable SEAs and LEAs to plan appropriate services designed to meet the child's unique needs.

3. *Instructional Planning Decisions.* Assessment data are routinely collected for the purpose of helping educational personnel plan instructional interventions for children. Specific efforts are made to identify an individual's educationally relevant strengths and weaknesses; to plan precisely what to teach and how to teach. Data collected during assessment serve as the basis for planning both long term educational goals and specific instructional objectives.
4. *Individual Pupil Evaluation.* Teachers, parents, and students themselves have a right and a need to know the extent to which pupils are progressing in their educational programs. Assessment data are provided that enable decision-makers to judge the extent to which progress is being made, both in the achievement of specific instructional objectives and in reference to a local or national sample of age- or grade-mates.
5. *Program Evaluation Decisions.* Data are collected for the purpose of evaluating the effectiveness of educational programs. Typically, this consists either of comparing the progress of students in two or more programs, or of looking at the extent to which pupils are attaining program objectives.

FACTORS TO BE CONSIDERED IN THINKING ABOUT IMPLEMENTATION OF THE PEP PROVISION

The Kind of Decision to be Made

It has been noted that assessment is the process of collecting data for the purpose of making decisions for and about students. Educators need different kinds of data to help make different kinds of decisions. Abuse in assessment can occur when educators fail to differentiate their decision-making procedures and the devices they use in light of the different kinds of decisions they make. This boils down essentially to the use of tests for purposes other than those for which they were designed. For example, norm-referenced, individually administered intelligence tests were originally designed to help us make classification and placement decisions. We witness, however, the routine practice of engaging in profile analyses of subject scores earned on norm-referenced tests in an effort to identify specific activity strengths and weaknesses, and the use of these data in instructional planning. Such a practice is without empirical support and may actually constitute abuse.

Acculturation

Any child's performance on a test is a reflection of past learning history in both formal and informal educational environments. As noted by Salvia and Ysseldyke (1978), acculturation is *the* most important characteristic in evaluating a child's performance on a test. To the extent that a child's acculturation differs from the acculturation of those on whom a test was standardized, norm-referenced decisions based upon test results may actually be both invalid and biased.

When norm-referenced tests are used to make decisions about students, assessors must examine the extent to which the student assessed is like those on whom the test was standardized. This is especially true when a child exhibits one or more specific handicapping conditions (e.g., deafness, blindness, or cerebral palsy).

Technical Adequacy

When assessment devices are used to make important decisions about pupils, it is imperative that those devices be technically adequate. Several factors must be considered. The first is again the issue of standardization. Scores earned on norm-referenced tests reflect the performance of the pupil relative to those on whom the test was standardized.

A second consideration is that of reliability. Reliability refers simply to the consistency with which a device measures a trait or set of behaviors. Reliability indices tell us how much error there is in any measure. In assessment, we are interested in obtaining results that adequately reflect student traits, characteristics, or behaviors; we want our results to be as free from error as is possible. Nunnally (1967) provides some standards regarding how reliable devices must be. He notes that when tests are used in experimentation, those tests must have reliabilities that exceed .50; and that when tests are used to make important decisions about pupils, those tests should have reliabilities that exceed .90.

It has been demonstrated (Salvia & Ysseldyke, 1978; Ysseldyke & Salvia, 1974) that very many of the norm-referenced devices used to make decisions about pupils lack the necessary reliability to be used in decision making. When unreliable tests are used to provide data for decision making, those decisions may be based more on error than on actual pupil characteristics.

The third issue regarding technical adequacy is the validity of the devices and procedures used to collect data about children. Tests must be valid for the particular purposes for which they are used. Assuming that tests measure what they purport to measure, when there is little empirical evidence to suggest they are valid, can lead to bias in decision-making.

Tests as Samples of Behavior

Tests are merely samples of behavior. In assessing a student's intelligence, for example, we do not directly measure intelligence. Rather, we observe the ways in which the student responds to sets of stimuli presented in a standardized format. Student performance leads us to infer degree of intelligence. "Intelligence", thus, is not an observable phenomenon, but an inferred construct.

Different tests sample different behaviors. Student performance on a test can only be viewed as a function of the kind(s) of behaviors sampled by the test. The greater the similarity between the kind(s) of behaviors sampled by a test and the kinds of behaviors to which predictions are being made, the lesser the degree of inference involved in assessment.

Bias in Decision-Making

It was noted earlier that the issue underlying the concerns expressed in the PEP provisions of P.L. 94-142 is bias in decision making. I believe that if educators suddenly had *the* fair test, there would still be considerable bias in decision-making. Recent research has demonstrated the extent to which naturally occurring characteristics act to bias the kinds of decisions made about pupils.

156

Ross and Salvia (1975) examined the extent to which students' physical attractiveness affects the kinds of decisions teachers make about those students. They sorted school pictures of third grade children into 10 piles, using a Q sort technique and asking raters to rank the pictures of students from those who were least physically attractive to those who were most physically attractive. The investigators selected pictures of children who had been rated most physically attractive and those rated least physically attractive, and affixed these to identical psychological reports. Reports included identical objective data regarding pupil intellect and prior achievement. Data provided were borderline data, data that could be used to support a label of either mentally retarded or normal. Four groups of classroom teachers were given the identical reports, but in one case the report included a picture of an attractive third grade boy, another report included a picture of an unattractive third grade boy, the third group received a report with a picture of an attractive third grade girl, while group four received a report with a picture of an unattractive third grade girl. Given identical objective information teachers reached different diagnostic decisions as a function of the physical attractiveness of the child. Attractiveness acted as a biasing factor in the kinds of diagnostic decisions reached.

Algozzine (1975) extended this research by looking at the extent to which pupils' physical attractiveness affected teacher-pupil interactions. He found that teachers interacted significantly less often and more negatively with unattractive than attractive pupils. Salvia, Algozzine and Sheare (1976) examined the grades that elementary teachers assign to attractive and unattractive pupils. They examined the cumulative records of children identified in the earlier study by Ross and Salvia as attractive and unattractive. They found no difference in the scores that these groups of students earned on intelligence tests and on measures of academic achievement. They found a one-grade-point difference in the grades assigned by teachers. Teachers were assigning higher grades to attractive than to unattractive students.

Further research on the extent to which naturally occuring characteristics affect decision-making was completed by Salvia and Podol (1975). They obtained a photograph of a child with a visible repaired cleft palate. They had the photograph retouched so the the repaired cleft was not visible. Two groups of speech therapists were given the same speech sample, and told they were to evaluate the speech of a child with a repaired cleft palate. One group was shown the picture of the child in which the repaired cleft was visible; the other group was shown the retouched photo. Significant differences were observed in the ratings of the same speech sample.

These and similar studies illustrate clearly bias in decision-making. Given objective data, decision-makers reached different conclusions as a function of diagnostically irrelevant pupil characteristics.

Bias Following Assessment

A series of recent investigations (Foster & Ysseldyke, 1972; Foster, Ysseldyke & Reese, 1975; Salvia, Clark & Ysseldyke, 1972; Ysseldyke & Foster, in press) has lead to concern regarding the extent to which the identification decisions we reach about pupils directly affect their later life opportunities. These investigations have examined the ways in which identification of children as handicapped causes teachers to view them differently and to misinterpret objective examples of their behavior.

Foster and Ysseldyke (1976) investigated the effects of deviancy-labels on teachers' expectations of child behavior and on their ability to evaluate child behavior objectively. One hundred elementary teachers were randomly assigned to one of four groups. Each group dealt with one label (emotionally disturbed, learning disabled, mentally retarded, normal), and each group participated in two separate treatment phases. During Phase I teachers identified behaviors they expected to be displayed by hypothetical children denominated by the label condition. Teachers indicated, for example, those behaviors they believed a "typical mentally retarded child" would demonstrate. During Phase II, each group viewed the same videotape of a normal fourth grade boy engaged in a variety of activities ranging from formal assessment to free play. After watching the videotape, teachers were asked to complete a checklist indicating the behaviors they had observed. Experimental conditions were identical across groups with one exception. Each group was told that the child was a member of a different category, that he was mentally retarded, emotionally disturbed, learning disabled, or normal.

Results of the investigation indicated that teachers (1) hold negative expectations for children to whom disability labels have been assigned, and (2) maintain these expectations even when confronted with normal behavior, behavior inconsistent with the stated label. Maintenance of this bias was sufficient to cause teachers to *misinterpret* actual child behavior, resulting in a halo effect. Results indicated that the label "educable mentally retarded" generated a greater degree of negative bias than did the labels "learning disabled" or "emotionally disturbed," although all three deviancy labels produced negative expectations and halo effects significantly different from those found under control conditions.

This body of research introduces another consideration into our thinking regarding bias. We need not only be concerned about how the decisions we make are biased, but at the same time about the effects of the decision-making process.

Summary

Nondiscriminatory assessment entails several factors in complex interaction. As I noted earlier, our real concern should be with bias in the decision-making process and with abuse in the use of assessment data to make decisions about students. Abuse can occur in many different ways. First, abuse can result from the use of tests for purposes other than those for which they were designed. It can also result from comparisons of students to others who differ systematically in several characteristics. Third, abuse occurs when technically inadequate tests are used to collect data about students. It also occurs when investigators go beyond their data to infer underlying pathology and infer or predict later academic difficulty. Bias on the basis of naturally occurring pupil characteristics occurs throughout the assessment process. Teachers differentially view objective child behavior when children are assigned deviancy labels.

FACTORS CONSIDERED IN DEVELOPING CRITERIA

No one set of criteria will serve universally to evaluate LEA implementation of the "protection in evaluation procedures" provisions of P.L. 94-142. School districts and states differ, both in the nature of the populations they serve and in the nature of the services they provide. Therefore, in developing criteria to evaluate implementation of the PEP provisions, several factors were taken into consideration.

Interrelationships of Stipulated Services

P.L.94-142 stipulates several different services — individualized educational programs, due process, placement in least restrictive environments, and protection in evaluation procedures. The PEP provisions are obviously related to the other three stipulated areas. As noted earlier, assessment is engaged in for the purpose of providing data that will help professionals make decisions about students. The different kinds of decisions were described.

Assessment is an integral component of the assessment-intervention process. Teachers routinely collect data about students prior to making decisions about the most appropriate instructional programs for them. It is required that assessment precede the making of educational placement decisions. In due process hearings parents and others are informed about and have a right to challenge the assessment data collected on their children. To the extent that abuse occurs in assessment (i.e., use of tests for purposes other than those for which they were designed, use of technically inadequate tests, etc.), biased decision-making can result.

159

Changes in Implementation Over Time

It is believed that LEAs will make steady progress toward implementation of the PEP provisions of P.L. 94-142. The criteria specified later in this paper are of the nature that such progress can be documented and demonstrated. Several criteria specify that LEAs routinely monitor their assessment procedures and evaulate the extent to which both the procedures and their effect are nondiscriminatory. Such monitoring should enable LEAs to spot areas of difficulty and to institute corrective efforts. It was thought that use of the criteria should lead an LEA from compliance with the letter of the law to eventual compliance with the spirit or intent of the law.

Contextual Influences

LEA contextual factors, such as its urban or rural environment, or the length of time that the LEA has been implementing state policies similar to those expressed in P.L. 94-142 are likely to influence implementation of the PEP provisions. The criteria developed later in this paper were developed so as to be useful and applicable in nearly all contexts. All LEAs, regardless of contextual factors, engage in decision-making and stand the chance of making biased or discriminatory decisions. While it is recognized that different LEAs make decisions about different kinds of constituencies, the criteria should apply across the board. This author has little regard for time considerations. If educational personnel are making important educational decisions about children, decisions that directly and significantly affect children's life opportunities, they should be using nondiscriminatory procedures in assessment.

Multiple Approaches

It is recognized that LEAs may employ different approaches or procedures in implementing the PEP provisions. The criteria developed later in this paper should apply, regardless of the specific approach used by an LEA.

Relationship of Criteria To Assessment Methodologies

The author has approached the task of developing criteria with one over-riding belief. The only assessment methodologies that should be employed in educational settings are those for which we have empirically demonstrated support. In very many instances today educational personnel collect data on children that are of little relevance to decision-making. For example, research

160

has repeatedly demonstrated the absence of support for the efficacy of perceptual-motor training designed to improve children's performance in reading. Yet, school systems continue to provide remedial programs characterized by perceptual-motor training and educators continue to believe that it is very important to assess children's modality preferences prior to prescribing instructional interventions (Arter & Jenkins, 1977; Ysseldyke, 1973, 1977). Clearly, a major consideration relevant to this paper is the extent to which educators engage in or believe in procedures for which there is little if any empirical support. The use of non-reliable tests (Ysseldyke & Salvia, 1974) to assign children to instructional programs for which there is no demonstrated support, represents abuse in assessment.

Definitions of Concepts

Special effort has been made, wherever it was believed necessary, to define the concepts used in the criteria.

SPECIFIC CRITERIA

Collection of Information

Referral

Assessment is a data collection process, and educational personnel routinely collect information on students and their families for the purpose of making educational decisions. Since 1974, when Congress passed Public Law 93-380, there has been an established set of procedures and regulations that schools must follow in data collection. Local education agencies should be following these guidelines, should have policies and procedures regarding the kinds of data they can collect, and clear guidelines regarding the obtaining of consent from parents in the data collection process.

The guidelines and regulations of Public Law 93-380, applicable to the collection of information on all pupils, are articulated further in Public Law 94-142, for it is very clear that specific procedures need to be followed when children are referred for consideration for special educational services. Implementation of the "Protection in Evaluation Procedures" (PEP) provisions of Public Law 94-142 requires that LEAs employ certain safeguards in the process of referring children for evaluation. The evaluation process begins with referral; failure to employ specific procedural safeguards can contribute to abuse in the assessment process.

The first consideration regarding referral should be for the kinds of behavior that warrant concern and referral. Children are typically referred for psychoeducational evaluation when the behavior they exhibit is sufficiently different from

normally expected behavior that a person in a position to do so becomes concerned and calls them to the attention of diagnostic personnel. To the extent that teachers do not have a good understanding of normal child development and a commensurate understanding of behaviors that are considered deviant, over- or under-referral of children can result. More importantly, failure to be aware of cultural differences and the extent to which specific behaviors are deemed appropriate and/or inappropriate in specific cultural environments can lead to over- or under-referral of students who are members of those specific cultural groups. Furthermore, the specific biases of individual teachers toward specific naturally-occurring pupil characteristics can lead teachers to over-refer or under-refer children who demonstrate those characteristics. Diagnostic personnel must have established procedures for monitoring the referral process in their LEAs and must systematically examine and evaluate that process.

Whenever students are referred for psychoeducational evaluation, educational personnel must inform parents of the specifics of the referral. Informed consent must be obtained from parents *prior* to the evaluation of their children. Parents must be told who referred the child, specifically those behaviors that are reasons for concern, and provided with objective documentation of the reasons for referral, in language they can understand. School personnel must no longer address evaluation in generalities like "We want to test your child to see if anything is wrong with her," and must not assume that parents will not understand the reasons why the child is being referred. Honesty in communication at this point will alleviate many potential difficulties later in the assessment and decision-making process. "Informed consent means that the parent (or pupil) is 'reasonably competent to understand the nature and consequences of his decision' (Russell Sage Foundation Conference Guidelines, 1969, p. 17)" (Salvia & Ysseldyke, 1978, p. 437).

Specific criteria to be used in judging the extent to which LEAs referral procedures are in compliance with the PEP provisions of PL 94-142 are as follows:

1. LEAs have established procedures regarding the kinds of data that can be collected on pupils.
2. LEA procedures regarding the kinds of data that can be collected on pupils are consistent with the guidelines and regulations of Public Law 93-380, the Family Educational Rights and Privacy Act.
3. Diagnostic personnel routinely meet with groups of teachers to provide training in the kinds of behaviors teachers should and should not be looking for in considering children for referral.
4. The LEA has a record of the number of children referred by individual teachers and regularly examines this record to ascertain the extent to which any one teacher has a history of over-referral of children from certain cultural groups or who demonstrate specific common characteristics.

5. In all evaluation procedures, diagnostic personnel carefully consider the extent to which cultural differences or naturally occurring pupil characteristics may have biased the decision to refer a child.
6. The LEA has established procedures for periodic evaluation of the extent to which cultural differences between teachers and children may lead to misinterpretation of child behavior and to unnecessary over-referral of children from specific cultural groups.
7. The LEA regularly examines its referral patterns to ascertain the extent to which naturally occurring pupil characteristics affect the decision to refer children for consideration for special services.
8. When children are referred for psychoeducational evaluation, the parents/guardians are informed that a referral has been made and are
   a. Told who made the referral.
   b. Told precisely why the referral was made.
   c. Provided with objective documentation of the reasons for referral in language they can understand.
9. Parents are regularly invited to participate in the assessment and decision-making process for their children.
10. Parents are informed of their right to examine relevant records with respect to the assessment of their child.

Participation in Decision Making

The second major area of concern relevant to the collection of information on children and their families concerns those who are to participate in the data collection and decision-making process. While this is obviously a function of the kind of decision to be made, it is readily apparent that 1) many different kinds of information are collected in the making of educational decisions, and 2) it is highly unlikely that any one person has either the time or the necessary competencies to engage in all phases of the data collection process.

Educators have spent considerable time debating the issue of who should assess children. The "my turf – your turf" debate has repeatedly been aired in both the professional literature and at professional meetings by special educators, remedial reading teachers, school psychologists, speech therapists, guidance counselors, social workers, and administrators. I am not as concerned with the issue of "who" assesses children as I am with the belief that children should be assessed only by those who have the necessary competencies to do so.

Clearly, the task of making important educational decisions for and about children is a significant enough task to demand both competence and multidisciplinary cooperation. Yet, I have no pat solution to the problem or issue of assuring that only "competent" persons make decisions about children. During the last decade we have witnessed repeated difficulty in defining competence in educational settings. When it comes to decision-making, the definition of competence is far more slippery. Many different kinds of

competence are required; the competencies necessary to engage in decision-making change as a function of the kind of decision to be made and the characteristics of the youngster about whom decisions are made. The competency issue is best solved in several inter-related ways. First, this matter requires considerable self-evaluation and individual responsibility. Educational personnel must be willing to recognize and admit their own limitations, to recognize that no one person is an "expert" in all areas, and to be willing to refer children to other personnel for certain parts of an evaluation. Second, psychoeducational decision-making must be completed by teams of personnel; teams in which each member is able to contribute both uniquely and collectively. Third, educational decisions must be subject to due process to insure that checks and balances are placed on decision-making procedures. In some instances, development of lists of competencies to be demonstrated by individual professionals may be helpful, though the absence of any "policing" mechanism usually leaves such endeavors ineffective.

Decisions regarding placement and/or planning of specific psychoeducational interventions are decisions that require the active participation of a multidisciplinary team. Individuals to be involved will necessarily differ both as a function of the setting and the particular child for whom a decision is being made. Differentiated staffing should characterize the decision-making process. Active participation refers specifically to *participation*, not simply involvement. When educational personnel make important decisions that directly and significantly affect students' life opportunities, it is assumed that they will actually have spent time observing or working directly with the child.

A third major factor becomes apparent when decisions are to be made regarding a student who is a member of a specific racial or cultural group. The phenomenon of cultural awareness must not be taken lightly by educational decision makers. One major contributor to past abuses in assessment has been the absence of decision making of a person or persons who had an adequate understanding of the child's culture. Ideally, minority group decision makers should participate in the decisions made about minority group children. At the very least, every placement and intervention planning team should include participants (other than the parents) who understand and are aware of the student's cultural background.

Finally, once again the role of parents and/or guardians is critical in the decision making process. LEAs must be able to document the fact that parents are active participants in the assessment and decision-making process. Parents should be consulted at the time of referral, should be treated as a valuable source of developmental data during the evaluation process, and should grant their informed consent to the decisions reached regarding their child. Schools are legally constituted extensions of the family; parents entrust schools with many responsibilities for their children. Only in those instances in which parental

desires and values are clearly believed contrary to the good of the child should the schools take legal action to overrule parental desires. Criteria specific to the issue of participation in decision-making include the following:

11. Classification and placement decisions are made by teams of personnel to include at least the following:
    a. A teacher who has taught the child.
    b. Teachers to whose room the child may be assigned.
    c. A certified school psychologist.
12. Educational personnel who administer tests to students have demonstrated competence in the correct administration, scoring, and interpretation of the tests they use. Demonstrated competence is typically required for certification or licensure as a school psychologist.
13. Diagnostic personnel readily recognize their limitations and routinely refer children to others with demonstrated competency in specific kinds of assessment.
14. Placemement teams demonstrate awareness of community resources and resource personnel who might assist the team in developing educational plans for children, and of resources that might provide related or other services needed by the child.
15. When placement decisions are made about children who are members of a minority culture, at least one member (other than the parent) of the decision-making team is a member of that minority culture.
16. Parents are regularly involved in the entire assessment and decision-making process.
17. The LEA has established procedures for periodic review of the decision-making process and has documented for each placement decision
    a. the participants in the decision-making process.
    b. the kinds of data collected and the reasons.
    c. the settings in which the child was observed, evaluated and the person responsible for data collection.
    d. the decision reached and primary factors considered in reaching the decision.

The Information Base

A third major area of consideration relative to the collection of information is the kind of information to be collected. It has been noted earlier that the major consideration governing the kind of decision to be made. The making of different kinds of educational decisions requires the collection of different kinds of information: LEAs must give evidence of engaging in differentiated data collection.

Numerous methodologies are employed in assessment, ranging from the collection of historical information by means of interview to the collection of current information regarding level of skill development by means of formal

165

psychometric appraisal. Salvia and Ysseldyke (1978) provided a matrix within which to view the sources of diagnostic information. Assessment procedures and methodologies differ as a function of both the kind of information collected and the time at which the information is collected. The matrix illustrating this is reproduced in Figure 1.

TYPE OF INFORMATION	TIME AT WHICH INFORMATION IS GATHERED: CURRENT	HISTORICAL
Observations	Frequency counts of occurrence of a particular behavior	Birth weight
		Anecdotal records
		Observations by last year's teacher
	Antecedents of behavior	
	Critical incidents	
Tests	Results of an intelligence test administered during the assessment	Results of a standardized achievement-test battery given at the end of last year
	Results of this week's spelling test given by the teacher	
Judgments	Parents' evaluations of how well the child gets along in family, neighborhood, etc.	Previous medical, psychological, or educational diagnoses
		Previous report cards
	Rating scales completed by teachers, social workers, etc.	Parents' recall of developmental history, of undiagnosed childhood illnesses, etc.
	Teacher's reason for referral	

FIGURE 1. Sources of diagnostic information, classified according to type of information and time at which the information is collected.

From J. Salvia & J. Ysseldyke (Eds.), *Assessment in Special and Remedial Education*. Boston, Mass.: Houghton Mifflin, 1978.

Diagnostic personnel too often approach an assessment by asking "What test is most appropriate for use with this child?" One does not have to venture far to hear debates about whether a child should be assessed using a norm-referenced test or a criterion-referenced test, a Stanford Binet or a Wechsler, a Stanford Achievement Test or the Iowa Tests of Basic Skills. There are specific criteria that should dictate the specific test to be used in assessing children. These are addressed below. However, two considerations are important at this time. First the process of assessment should begin by asking "What behaviors do I want to sample?" rather than "What test should I use?". Tests, observations, and interviews are merely samples of behavior. The kind of behavior to be sampled should be a function of both the reason for evaluation and the kind of decision to be made. Tests, or parts of tests, or specific test items are merely used in assessment to collect samples of pupil behavior.

P.L. 94-142 requires that multiple sources of data be used in decision-making. Two regulations specific to this point read as follows:

1. No *single* procedure shall be the sole criterion for determining an appropriate educational placement for a child.
2. The child is assessed in all areas related to the suspected disability, including, where appropriate, health, vision, hearing, social and emotional status, general intelligence, academic performance, communicative status, and motor abilities.

Models for looking at the kinds of data provided by assessment were described by Mercer and Ysseldyke (1977). They examined the Medical Model, Social System (Deviance) Model, Psychoeducational Process Model, Task Analysis Model, and Pluralistic Model in terms of a) definitions of abnormality, b) assumptions, c) characteristics of the models, d) characteristics of appropriate measures, e) ways in which scores are interpreted, f) the nature of treatments or interventions within each of the models, g) the extent to which each model has a racially or culturally discriminatory effect, and h) two incidental categories of information. Table 1 (from Mercer & Ysseldyke, 1977) lists the information. Mercer and Ysseldyke noted that each of the five assessment models, viewed separately, provides only a partial view of the child. Attempts to develop a nondiscriminatory diagnostic-intervention program should use a multimodal approach in which the child is viewed simultaneously from all five perspectives.

Not only must multiple models be used, but pupil behavior in multiple settings must be considered. In making decisions about individual children, educational personnel must consider the congruence between behaviors evidenced in different settings. When direct, naturalistic observations of pupil behavior and standardized test results are disparate, explanations ought to be sought before decisions are made. When multiple indices of pupil performance on psychometric devices sampling behavior from the same domain are incongruent,

TABLE 1

OUTLINE OF DIFFERENT ASSESSMENT MODELS

Elements of the Model	Medical Model	Social System (Deviance) Model	Psychoeducational Process Model	Task Analysis Model	Pluralistic Model
Definition of abnormal	Presence of biological symptoms of pathology.	Behavior that violates social expectations for specific role.	Psychoeducational process and/or ability deficits.	No formal definition of normal or abnormal. Each child is treated relative to himself and not in reference to a norm.	Poor performance when sociocultural bias controlled.
Assumptions	Symptoms caused by biological condition. Sociocultural background not relevant to diagnosis and treatment.	Multiple definitions of normal are role and system specific. Biological causation not assumed.	Academic difficulties are caused by underlying process and/or ability deficits. Children demonstrate ability strengths and weaknesses or abilities can be reliably and validly assessed. There are links between children's performance on tests and the relative effectiveness of different instructional programs.	Academic performance is a function of an interaction between enabling behaviors and the characteristics of the task. Children demonstrate skill development strengths and weaknesses. There is no need to deal with presumed causes of academic difficulties. There are skill hierarchies; development of complex skills is dependent upon adequate development of lower-level enabling behaviors.	Learning potential similar in all racial-cultural groups. Tests measure learning and are culturally biased.

(continued on next page)

Table 1—continued

Elements of the Model	Medical Model	Social System (Deviance) Model	Psychoeducational Process Model	Task Analysis Model	Pluralistic Model
Characteristics	Not culture bound. Deficit model.	Social system bound and role bound. Deficit and asset model.	Continuous model: degree of deficit. Evaluative: good development of psychoeducational processes necessary to academic success. Deficit model: norm-referenced. Disabilities or deficits are within the child. Deficits can exist unrecognized and undiagnosed. Completely culture bound.	Continuous model: degree of skill development. Bipolar with respect to specific skills. Evaluative: high level skill development better than low-level skill development. Subject matter referenced. Each child treated individually rather than in comparison to others. Idiographic. Based upon task analysis. Skill development influenced by sociocultural background. Completely culture bound.	Socioculturally bound. Asset model. Infers beyond test performance.
Characteristics of appropriate measures.	Measure biological symptoms. Validity determined by biological correlates.	Measure competence in social roles. Validity determined by correlates with group judgments.	Focus on deficits: measures of ability or psychoeducational process deficits. Norm-referenced assessment. Hypothetical internal determinants.	Focus on assessment of skills. Criterion-referenced assessment. Actual environmental determinants.	Culture-specific tests. Gain measures-test-retest. Pluralistic norms.

explanations must be sought. Scientific researchers routinely take into consideration sampling issues in their investigations. So too in assessment, which is essentially mini-experimentation designed to answer a diagnostician's hypotheses, multiple samples of behavior must be considered.

Salvia and Ysseldyke (1978) addressed in a very general sense the issue of differential data collection by noting the extent to which norm-referenced as opposed to criterion-referenced tests should be used in decision-making. The use of specific kinds of tests was viewed as a function of the kind of decision to be made. In screening, our primary concern is identification of the extent to which a student differs from others and further diagnostic appraisal is believed warranted. Such information is most readily and easily obtained by administration of norm-referenced tests. Results of pupil performance on technically adequate norm-referenced tests provide us a picture of the student's standing relative to others. Similarly, to remain accountable in the making of classification and placement decisions, educational administrators must be able to document the fact that a child is indeed sufficiently different from others that special educational services are warranted. Results of pupil performance on technically adequate norm-referenced devices are most useful in making classification and placement decisions.

When designing individual educational programs for students, teachers need and want to know specifically what to teach and how to teach. Such information is typically not transmitted by affording teachers the scores pupils earned on norm-referenced tests. Rather, teachers need to know specifically those skills that youngsters do and do not have, information readily obtained by administration of criterion-referenced measures. Information for use in evaluating individual pupil progress and relative to program evaluation can be obtained by means of both norm-referenced and criterion-referenced procedures. In the former, emphasis is on looking at pupil progress relative to that of others; in the latter evaluations consist of examining the extent to which pupils are attaining specific curricular objectives.

It is not my intent here to specify to a detail the data collection procedures (observation, interviewing, formal testing, informal testing, etc.) that should comprise an LEAs decision-making activities. Such decisions can best be made on an individual child basis. What is important to stress is the fact that *different* methodologies should characterize different decision-making needs, and LEAs should be able to demonstrate that they are engaged in differentiated assessment methodology.

Criteria for the selection of specific tests are easier to specify in considerably more detail, although the criteria result from a consideration of several factors in complex interaction. The specific tests selected for use in assessment are selected on the basis of an interaction between the kind of decision being made, the

acculturation of the pupil being assessed, and the technical adequacy of specific instruments. Yet, it is on this dimension that the greatest abuse in assessment has occurred.

Nearly every competent person who has been trained to assess children received education regarding the kinds of tests that could and could not be used. Yet, in practice, such considerations often are not observed. First, tests used in decision-making must be those that are designed for the purposes for which they are used. The most obvious abuse of this principle is observed in the profile analysis of data obtained from intelligence tests for use in planning individual educational programs for children. Intelligence tests are devices that were originally designed to assist decision-makers in classification and placement. They were not designed to be used in identification of specific diagnostic strengths and weaknesses for purposes of planning educational interventions. Furthermore, there has been little if any empirical support for the practice of using subtest profile analyses to plan specific programs for children (Mann, 1971; Ysseldyke, 1973).

Second, most assessors learned in their training that one of the primary considerations in selecting specific tests for use with students is one of acculturation. To the extent that norm-referenced tests are administered to a student for the purpose of providing information for use in decision-making and to the extent that the acculturation of the student assessed differs from the acculturation of those on whom the test was standardized, use of the device can contribute to abuse in assessment. Yet, it is readily apparent that this consideration is often overlooked in assessing pupils. It most certainly is overlooked in the assessment of specific kinds of handicapped pupils. Gerweck and Ysseldyke (1974), for example, responded to a survey conducted by Levine (1973), looking at the kinds of devices used to assess deaf students. In that survey, Levine reported that the most commonly used test in assessing deaf children was the Performance Scale of the Wechsler Intelligence Scale for Children. Those who use the device are violating a fundamental assumption in assessment by comparing deaf children to those whose acculturation has been radically different.

Third, the use of unreliable and invalid tests clearly contributes to abuse in decision-making. I have recommended that when tests are used to make placement decisions about students those tests should have reliabilities that exceed .90. This is the figure suggested by Nunnally (1967) and is as high as it is simply to reduce error in decision-making. This issue of high reliability causes special difficulty for assessors, because many norm referenced tests do not meet the specified criterion. It is my belief that only those tests that do have satisfactory reliability should be used in decision-making. Ysseldyke and Salvia (1974) published a list of the measures of specific processes and abilities often used in decision-making and concluded that nearly all such measures lacked the

necessary reliability to be used in decision-making. Others have argued that the use of unreliable tests is better than not using tests in decision-making. Ysseldyke and Salvia (1974) responded to this challenge by computing indexes of forecasting efficiency (coefficients of alienation) for each of the measures they studied. The use of unreliable tests did not significantly improve the prediction of pupil performance. Reliabilities and rates of improvement in prediction are listed in Table 2.

Salvia and Ysseldyke (1978) suggested alternatives to the use of unreliable norm-referenced tests in decision-making. One alternative consists of using estimated true scores rather than obtained scores in reporting test results and using them in decision-making. Any observed score is a function of the individuals true score plus error. The greater the amount of error in measurement, the greater the difference between a pupil's obtained score and her true score. Computing estimated true scores is one way of correcting for error in measurement. The formula to be used in computing estimated true scores is: $X' = \bar{X} + (rxx)(X - \bar{X})$.

Two examples of the procedure, one in which a reliable test is used, and one in which an unreliable test is used, may help clarify the reasons why I believe it is important to use estimated true scores.

Let's assume that Amy, a third grader, age 8-6, earns a Full Scale IQ of 85 on the Wechsler Intelligence Scale for Children (WISC-R), and a Psycholinguistic Quotient of 85 on the Illinois Test of Psycholinguistic Abilities (ITPA). Mean scores on both scales are 100, the reliability of the WISC-R Full Scale IQ is .95 for 8½ year-old student; the reliability of the ITPA PQ for eight-year-olds is .66. When we substitute these values into the above equation, we get estimated true scores of 86 and 90 (see Figure 2).

The principle is demonstrated that the estimated true score is a regression of the obtained score toward the mean. The lower the reliability of the obtained score, the greater the regression toward the mean.

A second alternative consists of restricting test use to devices having reliabilities greater than .90.

One other consideration is relevant to the issue of technical adequacy. Tests must have demonstrated validity for the purposes for which they are used. More than 10 years ago, a joint committee of the American Psychological Association, American Educational Research Association, and the National Council on Measurement in Education published a document entitled *Standards for Educational and Psychological Tests and Manuals* (APA, 1966). This document was revised in 1974 (APA, 1974) and re-titled *Standards for Educational and Psychological Tests*. The standards document stated that "A manual or research report should present the evidence of validity for each type of inference for

172

TABLE 2
RELIABILITIES OF FREQUENTLY USED TESTS

MEASURE	RELIABILITY
California Achievement Test (Subtest Reliabilities)	.76 - .97[a]
Iowa Test of Basic Skills (1974 edition)	none
Metropolitan Achievement Test	.84 - .96[c]
Stanford Achievement Test (1973 edition)	.65 - .97[a]
Gates-MacGinitie Reading Test	.88 - .96[c]
Peabody Individual Achievement Test	.42 - .94[b]
Wide Range Achievement Test	[c]
Gray Oral Reading Test	.97 - .98[d]
Gilmore Oral Reading Test	.53 - .94[d]
Gates-McKillop Reading Diagnostic Test	none
Durrell Analysis of Reading Difficulty	none
Stanford Diagnostic Reading Test (1976 edition)	.75 - .94[c]
Silent Reading Diagnostic Test	.85 - .97[c]
Diagnostic Reading Scales	.87 - .96[a]
Woodcock Reading Mastery Tests	.79 - .99[c]
Key Math	.39 - .90[a]
Stanford Diagnostic Mathematics Test	.84 - .97[a]
Stanford-Binet Intelligence Test	none
Wechsler Intelligence Scale for Children—Revised	
Verbal	.91 - .96[c]
Performance	.89 - .91[c]
Full Scale	.95 - .96[c]
Subtests	.62 - .92[c]
Wechsler Adult Intelligence Scale	
Verbal	.96[c]
Performance	.93 - .94[c]
Full Scale	.97[c]
Subtests	.60 - .95[c]
Wechsler Preschool and Primary	
Verbal	.93 - .95[c]
Performance	.91 - .95[c]
Full Scales	.96 - .97[c]
Subtests	.62 - .91[c]
McCarthy Scales of Children's Abilities	
Verbal	.86 - .92[c]
Perceptual-Performance	.75 - .90[c]
General Cognitive	.90 - .94[c]
Quantitative	.77 - .86[c]
Memory	.72 - .83[c]
Motor	.60 - .84[c]

173

which use of the test is recommended. If validity for some suggested interpretation has not been investigated, that fact should be made clear" (p. 31). The U.S. Office of Civil Rights expanded on this position in its publication of regulations relevant to section 504 of Public Law 93-112, the Rehabilitation Act of 1973. The Office of Civil Rights stated essentially that tests must have demonstrated validity for the purposes for which they are used. I am here articulating that position once again.

Abuse can and does occur in decision-making when invalid measures are used to provide data on students. In today's schools, invalid tests are very often used in decision-making. The clearest example of this can be illustrated by calling attention to a basic measurement principle all assessors learned during their training: Reliability is a necessary but not sufficient condition for validity. Given that many of the devices used to make decisions about pupils do not have adequate reliability, they cannot be said to be valid. The use of invalid measures to obtain data should cease.

The final issue relevant to the collection of information is a set of standards regarding the ways in which tests are administered. Inappropriate administration of tests can obviously contribute to inappropriate decision-making. Several considerations, most of them very obvious, are relevant to this point. Again, in test administration examiners must pay special attention to the acculturation of the individual assessed. The first step in test administration is selection of the behaviors to be sampled. Failure to consider the acculturation of the child and its relationship to the acculturation of those on whom a test was standardized can lead to serious errors in interpretation.

Local education agencies must also have ways of assuring that test administration is carried out by competent professionals. Competence includes skill in establishing rapport with children, as well as skill in correct administration, scoring, and interpretation of tests.

One other consideration is relevant to the administration of tests. The PEP provisions of Public Law 94-142 state that tests are to be administered in the child's native language or mode of communication. This provision creates special difficulty for LEA personnel. Clearly, if a child's native language differs from the language used in assessment, the potential for abuse in decision-making is indeed great. Yet, establishing children's native language in a culturally and linguistically diverse society is no easy task. Does one consider a child's native language to be the language spoken by the parents? If so, how does one go about assessing children who are, in fact, bilingual? Does one consider a child's native language to be the one which he/she first learned? Or does one consider the child's native language to be the one which he/she is now most fluent in? There is no readily apparent way to arrive at an answer to the dilema of assessing children in their native language. Furthermore, compliance with this provision is complicated by

two factors. First, there are very few technically adequate norm-referenced tests standardized in languages other than English. Second, the language of instruction in this nation's schools is English.

The provision of assessment of a child in his/her native language can be best addressed, I believe, by considering this issue as one part of the larger issue of acculturation. Children who have been reared in family environments where the principal language spoken is one other than English have clearly experienced an acculturation that differs from the acculturation of children on whom standardized tests were normed. Straight quantitative interpretation of test scores is obviously inappropriate. Rather, pupil performance must be looked at in light of the interaction between the acculturation of the individual assessed and the kinds of behaviors sampled by the devices and procedures used. I believe that at this time it is probably both impossible and meaningless to assess children in languages other than English. Compliance with this provision will necessarily have to come from increased intelligent use of tests and interpretation of pupil performance on tests.

The second part of the "native language" provision *can* be complied with at this time. It is possible to assess children using devices designed to be used with individuals who communicate in the same way. Deaf and hearing impaired children should be assessed using tests standardized on the deaf and hearing impaired. Blind children should be assessed using tests and procedures developed for use with the blind. Cerebral palsied youngsters should be assessed using tests devised for use with the cerebral palsied. A very common misuse of tests is witnessed in the use of tests standardized on non-sensorily-handicapped children to gather data for use in making decisions about sensorily handicapped individuals.

Specific criteria relevant to this section are as follows.

18. Local Education Agencies are able to document for every child about whom a placement decision is made, the following information:
    a. The primary language spoken in the child's home.
    b. Any unusual social and cultural customs of the child's family.
    c. The child's race.
    d. The extent to which the child may have a specific physical or sensory problem.
    3. That the child was observed in more than one environment (i.e., classroom, individual, play, home, etc.).

19. Educational personnel are able to document the fact that the assessment data they collect are relevant to the kind of decision they are making.
20. Educational personnel routinely observe children they assess in more than one setting (i.e., home, group instruction, individual instruction, play, etc.)

21. Observed scores are converted to estimated true scores prior to being interpreted.
22. Any time difference scores or deficit scores are used to identify children said to be handicapped, the reliability of those difference scores is computed and included in the report.
23. When norm-referenced devices are used to make decisions about children, diagnostic personnel are able to identify the extent to which the child assessed is like those on whom the test was standardized. Variables considered include the following:
    a. *Age*. Children of the same chronological age were included in the standardization group.
    b. *Grade*. Children of the same grade level were included in the standardization group.
    c. *Sex*. A representative sample of children of the same gender were included in the standardization group.
    d. *Acculturation of Parents*. The standardization group included a representative sample of children whose parents' acculturation was like that of the parents of the child assessed.
    e. *Geographic location*. There are children in the standardization sample who live in the same geographic region as the child assessed.
    f. *Date of Norms*. Norms for the tests used in decision-making are relatively current (within the last 15 years).
    g. *Special Population Characteristics*. If the child assessed has specific handicapping conditions (i.e., deafness, blindness, etc.), there are similar children in the standardization population.

24. When educational personnel use norm-referenced tests to make important decisions about children, they have evidence that the tests are valid for the purposes for which they are used.
25. Norm-referenced tests used to make important educational decisions about children have reliabilities that exceed .90.
26. Diagnostic personnel are able to state the reliabilities of the tests and subtests they use in decision making.
27. Children are always told why they are being assessed.
28. Diagnostic personnel are able to document the fact that in every assessment, the physical environment of the test setting has not adversely affected the child's performance. The following factors have been considered:
    a. Room temperature
    b. Noise
    c. Inadequate space
    d. Lighting
    e. Appropriateness of furnishings for the child's size.
29. Tests are administered according to the directions and procedures specified in the manual.
30. Adequate precautions were taken to insure that the examinee understood procedures and materials relevant to the test.

176

31. Test results are reported and interpreted within a range. Single scores are not used.
32. When placement decisions are made, diagnostic personnel give as much weight to adaptive behavior as they do to other data on the child.
33. Pupil behavior in multiple settings is sampled in the process of decision-making.
34. Decision-makers always gather more than one sample of behavior in any domain (intelligence, specific achievement, perceptual-motor, etc.)

Use of Assessment Information

A wise professor once said "There is nothing wrong with tests, it's the stupnagles that use them". Repeatedly we hear voiced the position that tests, in and of themselves are not bad, but that problems arise when they are misused. While such statements have appeal and a certain degree of validity, they must be modified in light of some of the points made earlier in this paper. There *are* numerous inadequately standardized norm-referenced tests that unfortunately are routinely used to make decisions about children. There *are* many technically inadequate tests that are routinely used to collect information on children. The previous section of this paper addressed concerns relevant to these points. This section develops criteria for evaluating the extent to which assessment information is *used* in a nondiscriminatory manner.

The Use of Test Information in Decision-Making

One of the most common abuses in assessment consists of using tests for purposes other than those for which they were designed. Earlier in this paper it was noted that educators use assessment data to make five different kinds of educational decisions. We need to be relatively specific regarding the kinds of information that legitimately ought to be collected in the process of making different kinds of decisions. In screening, our concern is for identification of those who are sufficiently different from others that additional assessment is believed warranted. Screening typically requires both observation and norm-referenced assessment. Similarly, placement and identification decisions are norm-referenced decisions. Educational personnel must be able to document the fact that a child is handicapped to provide services. Documentation usually consists of demonstrating that a child is sufficiently different from others. In planning instructional interventions, however, norm-referenced data are useful only to the extent that they are task-analyzed. Rather, as noted by Salvia and Ysseldyke (1978), teachers want and need to know specifically what to teach and how to teach. Information of this nature is best obtained by means of curriculum-based or criterion-referenced assessment. Evaluation decisions require collection of either norm-referenced (if one is attempting to compare pupil

177

performance to a national sample) or criterion-referenced (if one is ascertaining the extent to which students achieve objectives) information. Once again, the over-riding principle is that the kinds of information collected need to be a function of the kinds of decisions one is making.

35. In planning instructional interventions, diagnostic personnel place primary emphasis on data obtained by means of curriculum-based assessment.
36. LEAs routinely examine the inter-relationships between the kinds of data they collect and the kinds of decisions they make.

Bias in and Following Assessment

When assessment data are collected for the purpose of making decisions about pupils, and when those data are collected using assessment devices and procedures that are less than technically adequate, several subjective factors can influence the decision-making process. Earlier, it was indicated that much recent research demonstrates the fact that many naturally occurring pupil characteristics can and do act to bias the kinds of decisions made about pupils. LEAs must take steps to alleviate the extent to which biased decisions are made. This is probably the most difficult area in which to specify criteria. LEAs must provide training in decision-making and must document, to the extent possible, those factors considered primary in decision making. Mechanisms must be available for individual decision-makers to routinely examine the extent to which the decisions they make are free of bias.

37. LEAs provide training in objective decision-making for their decision-making personnel.
38. LEAs have established procedures for examining the extent to which the decisions they make are biased by subjective pupil characteristics.

Documentation of Decision-Making

It is critical that LEA personnel have a good understanding of the ways in which they make decisions about pupils. My own experience working with LEA personnel indicates that there are many varied opinions regarding the ways in which decisions are made, but little data to support those opinions. Many LEA personnel indicate that pupil performance on psychometric measures is the critical factor influencing decisions made about those pupils. Others state that pupil performance on tests is simply treated as one source of information in the decision-making process. Still others maintain that test-based information is seldom used in decision-making, that they go considerably beyond test scores to consider the "whole child". LEA personnel should be collecting data on and maintaining records regarding the decision-making process. Decision-making personnel should be required to provide a very brief rationale for each decision made, and routinely the decision-making process should be studied.

39. LEA decision-making personnel maintain records or other forms of documentation regarding those factors considered primary in decision-making.

40. LEA diagnostic personnel routinely evaluate the decision-making process and are able to identify the factors that are regularly considered primary in the placement, instructional planning, and evaluation decisions they make.

Reporting Scores in Ranges

Very many psychometric instruments routinely used to gather information for use in decision-making are technically inadequate. Reliance on scores earned on technically inadequate tests can contribute tremendously to abuse in decision-making. It was noted earlier that some sources of error in assessment can be reduced by converting obtained scores to estimated true scores. We can further reduce error in interpretation by reporting pupil performance in ranges rather than in single scores. I am recommending that in interpreting pupil performance on norm-referenced tests, diagnostic personnel first convert obtained scores to estimated true scores and then construct either symmetrical or assymmetrical confidence intervals around those estimated true scores. Procedures for doing so are outlined in Salvia and Ysseldyke (1978, pp. 85-88).

41. When reporting pupil performance on norm-referenced psychometric devices, diagnostic personnel first convert obtained scores to estimated true scores, and report performance only in terms of a range.

FIGURE 2
COMPUTATION OF ESTIMATED TRUE SCORES ON TWO DEVICES DIFFERING IN DEGREE OF RELIABILITY

$X' = \bar{X} + (r_{xx})(X - \bar{X})$	$X' = \bar{X} + (r_{xx})(X - \bar{X})$
$X' = 100 + (.95)(85-100)$	$X' = 100 + (.66)(85-100)$
$X' = 100 + (.95)(-15)$	$X' = 100 + (.66)(-15)$
$X' = 100 - 14.25$	$X' = 100 - 9.90$
$X' = 85.75$	$X' = 90.1$
$X' = 86$	$X' = 90$

REFERENCES

Algozzine, R. A. Attractiveness as a biasing factor in teacher-pupil interactions. University Park, PA: Unpublished doctoral dissertation, 1975.

American Psychological Association. *Standards for educational and psychological tests*. Washington, D.C.: APA, 1972.

Angoff, W. H. & Ford, S. F. *Item-race interaction on a test of scholastic aptitude*. Princeton, N.J.: Educational Testing Service, 1971. (ERIC Document Reproduction Service No. ED 058 279).

Angoff, W. H., & Sharon, A. T. The evaluation of differences in test performance of two or more groups. *Educational and Psychological Measurement*, 1974, *34*, (4), 807-816.

Arter, J. A., & Jenkins, J. R. Examining the benefits and prevalence of modality considerations in special education. *Journal of Special Education*, 1977, *11*, 281-298.

Ballard, J., & Zettel, J. Public law 94-142 and Sec. 504: What they say about rights and protections. *Exceptional Children*, 1977, *44*, 177-185.

Barnes, E. *IQ testing and minority school children: Imperatives for change*. Storrs, Conn.: Connecticut University, Leadership Institute-Teacher Education Early Childhood, 1973. (ERIC Document Reproduction Service No. ED 078 006)

Barnes, E. J. IQ testing and minority school children: Imperatives for change. *Journal of Non-White Concerns*, 1973, *2*(1), 4-19.

Barnes, N. R., Grill, J. J., & Bryen, D. N. Language characteristics of Black children: Implications for assessment. *Journal of School Psychology*, 1973, *11*(4), 351-364.

Bayley, N. Comparisons of mental and motor test scores for ages 1 - 15 months by sex, birth order, race, geographical locations and education of parents. *Child Development*, 1965, *36*, 379-411.

Bennett, V. C., & Bardon, J. I. Law, professional practice, and professional organizations: Where do we go from here. *Journal of School Psychology*, 1975, *13*(4), 349-358.

Bereiter, C. The future of individual differences. *Harvard Educational Review*, 1969, *39*, 162-170.

Berry, G. L., & Lopez, C. A. Testing programs and the Spanish-speaking child: Assessment guidelines for school counselors. *School Counselor*, 1977, *24*,(4), 261-269.

Bersoff, D. N., & Ysseldyke, J. E. Nondiscriminatory assessment: The law, litigation, and implications for the assessment of learning disabled children. Invited Address: Association for Children with Learning Disabilities Annual International Conference, 1977.

Bijou, S. W. Environment and intelligence: A behavioral analysis. In R. Cancro (Ed.), *Intelligence: Genetic and environmental contributions*. New York: Grune & Stratton, 1971.

Bloom, B. S. Stability and change in human characteristics. New York: Wiley, 1964.

Bodmer, W. F. & Cavalli-Sforza, L. L. Intelligence and race. *Scientific American*, 1970, *223*, 19-29.

Boone, J. A., & Adesso, V. J. Racial differences on a black intelligence test. *Journal of Negro Education*, 1974, *63*(4), 429-436.

Breland, H. M. *An investigation of cross-cultural stability in mental test items*. Paper presented at the annual meeting of the American Educational Research Association, Chicago, April 1974.

Breland, H., et al. *The cross-cultural stability of mental test items. An investigation of response patterns for ten socio-cultural groups* (Final Report). Princeton, NJ: Educational Testing Service, February, 1974. (ERIC Document Reproduction Service No. ED 137 370)

Breland, H. M., et al. *Cross-cultural stability of test items: An investigation of response patterns for ten socio-cultural groups with exploration of an index of cross-cultural stability* (Final Report). Princeton, N.J.: Educational Testing Service, December 1973. (ERIC Document Reproduction Service No. ED 115 682).

Bryen, D. N. Special education and the linguistically different child. *Exceptional Children*, 1974, *40*(8), 589-599.

Burt, C. Intelligence and achievement. *Mensa Register*, 1967, 1-2.

Butcher, J. *Human intelligence: Its nature and assessment*. London: Methuen, 1968.

Butler, O. T., Coursey, R. D., & Gatz, M. Comparison of the Bender Gestalt Test for both Black and white brain-damaged patients using two scoring systems. *Journal of Consulting and Clinical Psychology*, 1976, *44*(2), 280-285.

Campbell, J. Differential response for female and male law students on the Strong-Campbell Interest Inventory: The question of separate sex norms. *Journal of Counseling Psychology*, 1976, *23*(2), 130-135.

Cattell, R. B. Research designs in psychological genetics with special reference to the multiple variance analysis method. *American Journal of Human Genetics*, 1953, *5*, 76-93.

Cattell, R. B. Theory of fluid and crystallized intelligence: A critical experiment. *Journal of Educational Psychology*, 1963, *54*, 1-22.

Cattell, R. B. *Beyondism: The morality of science*. New York: Pergamon, 1971.

Cervantes, R. A. *Problems and alternatives in testing Mexican American students*. Paper presented at the annual meeting of the American Educational Research Association, Chicago, April 1974.

Cicchetti, D. V. A computer program for assessing the reliability and systematic bias of individual measurements. *Educational and Psychological Measurement*, 1976, *36*(3), 761-764.

Cleary, T. A. Test bias: Prediction of grades of Negro and White students in integrated colleges. *Journal of Educational Measurement*, 1968, *5*, 115-124.

Cole, N. S. Bias in selection. *Journal of Educational Measurement*, 1973, *10*, 237-255.

Coordinating Office for Regional Resource Centers. *With bias toward none: Proceedings of a National Planning Conference on Nondiscriminatory Assessment*. Lexington, KY.: Coordinating Office for Regional Resource Centers, 1977. (ERIC Document Reproduction Service No. ED 138 028)

Cronbach, L. J. Heredity, environment, and educational policy. *Harvard Educational Review*, 1969, *39*, 190-199.

Cronbach, L. J. Five decades of public controversy over mental testing. *American Psychologist*, 1975, *30*(1), 1-14.

Cronbach, L. J. Equity in selection: Where psychometrics and political philosophy meet. *Journal of Educational Measurement*, 1976, *13*(1), 31-41.

Crow, J. F. Genetic theories and influences: Comments on the value of diversity. *Harvard Educational Review*, 1969, *39*, 153-161.

Darlington, R. B. Another look at "cultural-fairness". *Journal of Educational Measurement*, 1971, *8*, 71-82.

DeGeorge, G. P. *Guidelines for selecting tests for use in bilingual/bicultural education programs.* Paper presented at the Matsol Spring Conference, 1975. (ERIC Document Reproduction Service No. ED 108 529)

Diamond, E. E. Minimizing sex bias in testing. *Measurement and Evaluation in Guidance*, 1976, *9*(1), 28-33.

Dreger, R. M. & Miller, K. S. Comparative psychological studies of Negroes and whites in the United States. *Psychological Bulletin Monograph Supplement*, 1968, *70*, No. 3, Part 2.

Durovic, J. J. *Test bias: An objective definition for test items.* Paper presented at the annual meeting of the Northeastern Educational Research Association, Ellenville, N.Y., October 1975.

Dwyer, C. A. *Test content in mathematics and science: The consideration of sex.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco, April 1976.

Ebel, R. L. Educational tests: Valid? biased? useful. *Phi Delta Kappan*, 1975, *57*(2), 83-88.

Eckland, B. K. Social class structure and the genetic basis of intelligence. In R. Cancro (Ed.), *Intelligence: Genetic and environmental influences*. New York: Grune & Stratton, 1971.

Eells, K., Davis, A., Havighurst, R. J., Herrick, R., & Cronbach, L. J. *Intelligence and cultural differences*. Chicago: University of Chicago Press, 1951.

Einhorn, H. J., & Bass, A. R. Methodological considerations relevant to discrimination in employment testing. *Psychological Bulletin*, 1971, *75*, 261-269.

Elkind, D. Piagetian and psychometric conceptions of intelligence. *Harvard Educational Review*, 1969, *39*, 171-189.

Epps, E. G. Race intelligence, and learning: Some consequences of the misuse of test results. *Phylon*, 1973, *34*(2), 153-159.

Erlenmeyer-Kimling, L. & Jarvik, L. F. Genetics and intelligence: A review. *Science*, 1963, *142*, 1477-1479.

Evans, H. I., & Sperekas, N. B. Reply to "Sex differences in adaptive styles." *Journal of Genetic Psychology*, 1975, *127*(2), 317-318.

Faggen-Steckler, J., et al. A quantitative method for measuring sex "bias" in standardized tests. *Journal of Educational Measurement*, 1974, *11*(3), 151-161.

Fishbein, R. L. *An investigation of the fairness of the items of a test battery*. Paper presented at the annual meeting of the National Council on Measurement in Education, Washington, D.C., April 1975.

Fitzgibbon, T. J. *Evaluation in the inner city*. New York: Harcourt Brace Jovanovich, 1971.

Flaugher, R. L. *Bias in testing: A review and discussion* (TM Report No. 36) Princeton, N.J.: ERIC Clearinghouse on Tests, Measurement, and Evaluation, 1974. (ERIC Document Reproduction Service No. ED 099 431)

Foster, G. G., & Ysseldyke, J. E. Expectancy and halo effects as a result of artificially induced teacher bias. *Contemporary Educational Psychology*, 1976, *1*, 37-45.

Foster, G. G., Ysseldyke, J. E. & Reese, J. I wouldn't have seen it if I hadn't believed it. *Exceptional Children*, 1975, *41*, 469-473.

Franklin, A. J. *The testing dilemma for minorities*. Paper presented at the public hearings on statewide testing and evaluation, Albany, New York, October 1974. (ERIC Document Reproduction Service No. ED 103557)

Frazer, W. G., et al. Bias in prediction: A test of three models with elementary school children. *Journal of Educational Psychology*, 1975, *67*(4), 490-494.

Gerweck, S., & Ysseldyke, J. E. Limitations of current psychological practices for the intellectual assessment of the hearing impaired: A response to the Levine survey. *Volta Review*, 1974, *77*, 243-248.

Gilmore, G., et al. The Bender Gestalt and the Mexican American student: A report. *Psychology in the Schools*, 1975, *12*(2), 172-175.

Ginsberg, B. E., & Laughlin, W. S. Race and intelligence: What do we really know? In R. Cancro (Ed.), *Intelligence: Genetic and Environmental Contributions*. New York: Grune & Stratton, 1971.

Gold, M. G., & Bruno, J. F. The judicial-legal definition of discrimination in testing. *Education and Urban Society*, 1975, *8*(1), 7-18.

Goldman, R. D., & Hewitt, B. N. Predicting the success of Black, Chicano, Oriental and white college students. *Journal of Educational Measurement*, 1976, *13*(2), 107-117.

Gordon, E. W. Methodological problems and pseudoissues in the nature-nurture controversy. In R. Cancro (Ed.), *Intelligence: Genetic and environmental contributions*. New York: Grune & Stratton, 1971.

Green, D. R. *Biased tests*. Monterey, Calif.: CTB/McGraw Hill, 1971.

Green, D. R. *Racial and ethnic bias in test construction*. Monterey, Calif.: CTB/McGraw-Hill, 1971.

Green, D. R. & Roudabush, G. E. An investigation of bias in a criterion-referenced test. (ERIC Document Reproduction Service No. ED 113 379)

Green, R. L., et al. *Standardized achievement testing: Some implications for the lives of children*. Paper presented at the National Institute of Education Test Bias Conference, Washington, D.C., December 1975.

Green, W. How testing harms children. *South Today*, 1973, *4*(8), 6-7.

Greenfield, P. M. Goal as environmental variable in the development of intelligence. In R. Cancro (Ed.), *Intelligence: Genetic and environmental contributions*. New York: Grune & Stratton, 1971.

Guilford, J. P. *The nature of human abilities*. New York: McGraw-Hill, 1967.

Guilliams, C. T. *Item analyses of American Indian and Chicano responses on the vocabulary scales of the Stanford-Binet LM and Wechsler batteries* (Final Report). Washington, D. C.: National Institute of Education, January 1976. (ERIC Document Reproduction Service No. ED 111 878)

Guion, R. Employment tests and discriminatory hiring. *Industrial Relations*, 1966, *5*, 20-37.

Hambleton, R. K., et al. *Developments in latent trait theory: A review of*

models, technical issues, and applications. Paper presented at a joint meeting of the National Council on Measurement in Education and the American Educational Research Association, New York, April 1977.

Harmon, L. W. Sexual bias in interest measurement. *Measurement and Evaluation in Guidance*, 1973, *5*(4), 496-501.

Hartlage, L. C., & Lucas, T. L. Differential correlates of Bender-Gestalt and Beery Visual Motor Integration Test for Black and for white children. *Perceptual and Motor Skills*, 1976, *43*(4), 1039-1042.

Hannessy, J. J. & Merrifield, P. R. A comparison of the factor structures of mental abilities in four ethnic groups. *Journal of Educational Psychology*, 1976, *68*(6), 754-759.

Hirsch, J. Behavior-genetic analysis and its biosocial consequences. In R. Cancro (Ed.), *Intelligence: Genetic and environmental contributions.* New York: Grune & Stratton, 1971.

Holland, J. L., et al. Sex differences, Hem revisions, validity, and the self-directed search. *Measurement and Evaluation in Guidance*, 1976, *8*(4), 224-228.

Hoepfner, R., & Strickland, G. P. *Investigating test bias.* Los Angeles: California University, Center for the Study of Evaluation, 1972. (ERIC Document Reproduction Service No. ED 066 443)

Humphreys, L. G. *Fairness for individuals and fairness for selection: Some basic considerations.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, 1973.

Hunt, J. McV. *Intelligence and experience.* New York: Ronald, 1961.

Hunt, J. McV. Has compensatory education failed? Has it been attempted? *Harvard Educational Review*, 1969, *39*, 130-152.

Hunt, J. McV. *Psychological assessment in education and social class.* Paper presented at the annual Missouri Conference on the Legal and Educational Consequences of the Intelligence Testing Movements: Handicapped Children and Minority Group Children, University of Missouri – Columbia, April 1972. (ERIC Document Reproduction Service No. ED 077 943)

Hunt, J. McV. & Kirk, G. E. Social aspects of intelligence: Evidence and issues. In R. Cancro (Ed.), *Intelligence: Genetic and environmental contributions.* New York: Grune & Stratton, 1971.

Hunter, J. E., & Schmidt, F. L. Critical analysis of the statistical and ethical implications of various definitions of test bias. *Psychological Bulletin*, 1976, *83*(6), 1053-1071.

Hunter, J. E., & Schmidt, F. L. *Fairness of selection tests: A critical analysis*. (Professional Series No. 76-5). Washington, D.C.: Personnel Measurement Research and Development Center, September 1976. (ERIC Document Reproduction Service No. ED 137 351)

Jensen, A. R. Estimation of the limits of heritability of traits by comparison of monozygotic and dizygotic twins. *Proceedings of the National Academy of Sciences*, 1967, *58*, 149-156.

Jensen, A. R. Social class, race, and genetics: Implications for education. *American Educational Research Journal*, 1968a, *5*, 1-42.

Jensen, A. R. Patterns of mental ability and socio-economic status. *Proceedings of the National Academy of Sciences*, 1968b, *60*, 1330-1337.

Jensen, A. R. How much can we boost I.Q. and scholastic achievement? *Harvard Educational Review*, 1969a, *39*, 1-123.

Linn, R. L., & Werts, C. E. Considerations for studies of test bias. *Journal of Educational Measurement*, 1971, *8*, 1-4.

Lockheed-Katz, M. *Sex bias in educational testing: A sociologist's perspective* (Research memorandum No. 74-13). Princeton, N.J.: Educational Testing Service, 1974. (ERIC Document Reproduction Service No. ED 098 262)

Long, P. A., & Anthony, J. J. The measurement of mental retardation by a culture-specific test. *Psychology in the Schools*, 1974, *11*(3), 310-312.

Lord, F. M. *A study of item bias using characteristic curve theory*. New York, N.Y.: College Entrance Examination Board, August 1977. (ERIC Document Reproduction Service No. ED 137 486)

Matluck, J. H., & Mace, B. J. Language characteristics of Mexican American children: Implications for assessment. *Journal of School Psychology*, 1973, *11*(4), 365-386.

Matluck, J. H., & Mace-Matluck, B. J. Language and culture in the multi-ethnic community: Spoken-language assessment. *Modern Language Journal*, 1975, *59*(5-6), 250-255.

187

Matuzek, P. A., & Oakland, T. D. *A factor analysis of several reading readiness measures for different socioeconomic and ethnic groups.* Paper presented at the annual meeting of the American Educational Research Association, Chicago, April 1972.

McClelland, D. C. Testing for competence rather than for intelligence. *American Psychologist*, 1973, *28*(1), 1-14.

McNemar, Q. On so-called test bias. *American Psychologist*, 1975, *30*(8), 848-851.

McNeil, N. D. An investigation of performance differences of urban high school students by race, sex, and grade on a measure of cultural homogenity and intelligence. (Ann Arbor, Mich.: University Microfilms No. 75-27, 176)

McNemar, Q. On so-called test bias. *American Psychologist*, 1975, *30*(8), 848-851.

Meeker, M., & Meeker, R. Strategies for assessing intellectual patterns in Black, Anglo, and Mexican-American boys — or any other children — and implications for education. (ERIC Document Reproduction Service No. ED 084 909)

Mercer, J. R. *Sociocultural factors in the educational evaluation of Black and Chicano children.* Paper presented at the tenth annual conference on civil rights educators and students, NEA, Washington, D.C., February, 1972. (ERIC Document Reproduction Service No. ED 062 462)

Mercer, J. R. *The origins and development of the pluralistic assessment project.* Sacramento, Calif.: California State Department of Mental Hygiene, Bureau of Research, 1972. (ERIC Document Reproduction Service No. ED 062 461)

Mercer, J. Personal Communication, 1977.

Mercer, J., & Lewis, J. *System of multicultural pluralistic assessment.* New York: Psychological Corporation, 1978.

Mercer, J., & Ysseldyke, J. E. Designing diagnostic-intervention programs. In T. Oakland (Ed.), *Psychological and educational assessment of minority children.* New York: Brunner-Mazel, 1977.

Marz, W. R. *Estimating bias in test items utilizing principal components analysis and the general linear solution.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco, April 1976.

Merrifield, P. R. Using measured intelligence intelligently. In R. Cancro (Ed.), *Intelligence: Genetic and environmental influences*. New York: Grune & Stratton, 1971.

Messick, S., & Anderson, S. *Educational testing, individual development, and social responsibility*. Princeton, N.J.: Educational Testing Service, 1970. (ERIC Document Reproduction Service No. ED 047 003)

Meyers, E. D. *The revised WISC: Does it serve inner city children*. Paper presented at the annual convention of the American Orthopsychiatric Association, Atlanta, March 1976.

Mizelle, R. M. Cultural diversity: Theory and technology. *High School Journal*, 1976, *60*(2), 57-64.

Moran, R. E. *Observations and recommendations on the Puerto Rican version of the Wechsler Intelligence Scale for Children*. Rio Piedras, Puerto Rico: Puerto Rico Universtiy, College of Education, 1974. (ERIC Document Reproduction Service No. ED 088 932)

Mukherjee, A. K., et al. *Measurement of intellectual potential Mexican-American school-age children*. Austin, Tex.: Texas Education Agency, Texas State Department of Mental Health, June 1976. (ERIC Document Reproduction Service No. ED 138 034)

Nathanson, D. E. Placement tests and the linguistically different: Discrimination in the guise of legality. *Negro Educational Review*, 1975, *26*(1), 52-59.

National Association for the Advancement of Colored People. *NAACP report on minority testing*. New York: College Entrance Examination Board, May 1976. (ERIC Document Reproduction Service No. ED 128 535)

Nafte, M. C. *School testing, grouping and the law*. Paper presented at the annual meeting of the National Organization on Legal Problems of Education, Colorado Springs, November 1975.

Neal, A. W. Analysis of responses to items on the Peabody Picture Vocabulary Test according to race and sex. *Dissertation Abstracts International*, 1975, *36*(2-A), 789-790.

Newland, T. E. Assumptions underlying psychological testing. *Journal of School Psychology*, 1973, *11*(4), 316-322.

Nunnally, J. *Psychometric theory*. New York: McGraw-Hill, 1967.

189

Ortega, F. Special education placement and Mexican Americans. *El Grito*, 1971, *4*(4), 29-35.

Peck, R. L. A comparative analysis of the performances of Indian and White children from north central Montana on the Wechsler Intelligence Scale for Children. *Dissertation Abstracts International*, 1973, *33*(8-A), 4097.

Petersen, N. S. & Novick, M. R. *An evaluation of some models for test bias* (Technical Bulletin No. 23). Iowa City, Iowa: American College Testing Program, Research and Development Division, September 1974. (ERIC Document Reproduction Service No. ED 128 372)

Petersen, N.S., & Novick, M. R. An evaluation of some models for culture-fair selection. *Journal of Educational Measurement*, 1976, *13*(1), 3-29.

Pfeifer, C. M., Jr., & Sedlacek, W. E. The validity of academic predictors for black and white students at a predominantly white university. *Journal of Educational Measurement*, 1971, *8*(4), 253-261.

Pine, S. M., & Weiss, D. J. Effects of item characteristics on test fairness (research Report No. 76-5). Minneapolis, Minn.: University of Minnesota, Department of Psychology, December 1976. (ERIC Document Reproduction Service No. ED 134 612)

Poole, R. L. Evaluating and victimizing elementary school children. *Education*, 1976, *97*(2), 115-120.

Prediger, D. J., & Hanson, G. R. *Evidence related to issues of sex bias in interest inventories*. Paper presented at the 84th annual convention of the American Psychological Association, Washington, D. C., September 1976.

Ratusnik, D. L., Koenigsknecht, R. A. *Drawing test performance of Black and White preschoolers as a function of biracial testing*. Paper presented at the annual meeting of the American Educational Research Association, Washington, D.C., April 1975.

Ratteray, J. D. *The testing of cultural groups. A paradigmatic analysis of the literature on testing and a proposition*. Santa Monica, Calif.: Rand Corporation, November 1974. (ERIC Document Reproduction Service No. ED 113 371)

Reschly, D. J., et al. *Analysis of different concepts of cultural fairness using WISC-R and MAT scores from four ethnic groups*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, April 1976.

Rincon, E. L. Comparison of the cultural bias of the KIT: EXP with the WISC using Spanish surname children differing in language spoken. *Educational and Psychological Measurement*, 1976, *36*(4), 1037-1041.

Ross, M. & Salvia, J. Attractiveness as a biasing factor in teacher judgements. *American Journal of Mental Deficiency*, 1975, *80*, 96-98.

Rudner, L. M. *An approach to biased item identification using latent trait measurement theory*. Paper presented at the annual meeting of the American Educational Research Association, New York, April 1977.

Salvia, J. Algozzine, R., & Sheare, J. Attractiveness and school achievement. *Journal of School Psychology*, 1976.

Salvia, J., Sheare, J., & Algozzine, R. Facial attractiveness and personal-social adjustment. *Journal of Abnormal Child Psychology*, 1975, *3*, 171-178.

Salvia, J., & Ysseldyke, J. E. *Assessment in special and remedial education*. Boston: Houghton-Mifflin, 1978.

Samuda, R. J. *Racial discrimination through mental testing: A social critic's point of view* (IRCD Bulletin No. 42). New York: Columbia University, ERIC Clearinghouse on the Urban Disadvantaged, 1973. (ERIC Document Reproduction Service No. ED 092 648)

Scales, A. M., & Smith, G. S. Strategies for humanizing the testing of minorities. *Negro Educational Review*, 1974, *25*(4), 174-180.

Scheyneman, J. *Validating a procedure for assessing bias in test items in the absence of an outside criterion*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, April 1976.

Sedlacek, W. E. *Recent developments in test bias research* (Research Report No. 2-76). College Park, MD.: University of Maryland, Cultural Study Center, January 1977. (ERIC Document Reproduction Service No. ED 127 532)

Sharf, J. C. Fair employment implication for HRD: The case of Washington vs. Davis. *Training and Development Journal*, 1977, *31*(2), 16-18, 20-21.

Simon, A. J. & Joiner, L. M. *Adapting the Peabody Picture Vocabulary Test for use with Mexican children*. Paper presented at the annual meeting of the American Educational Research Association, Chicago, April 1974.

191

Smith, M. W. Alfred Binet's remarkable questions: A cross-national and cross-temporal analysis of the cultural biases built into the Stanford-Binet Intelligence Scale and other Binet tests. *Genetic Psychology Monographs*, 1974, *89*(2), 307-334.

Southwest Regional Resource Center. *Unbiased assessment: Guidelines, procedures, and forms for the SEA's implementation of Public Law 94-142*. Salt Lake City, Utah: Southwest Regional Resource Center, January 1977. (ERIC Document Reproduction Service No. ED 138 024)

Strassberg-Rosenberg, B., & Donlon, T. F. *Content influences on sex differences in performance on aptitude tests*. Paper presented at the annual meeting of the National Council on Measurement in Education, Washington, D.C., April 1975.

Temp, G. Validity of the SAT for blacks and whites in thirteen integrated institutions. *Journal of Educational Measurement*, 1971, *8*(4), 245-251.

Thorndike, R. L. Concepts of culture-fairness. *Journal of Educational Measurement*, 1971, *8*, 63-70.

Tinsley, H. E. A., & Dawis, R. V. *The equivalence of semantic and figural test presentation of the same items*. Minneapolis, Minn.: Minnesota University, Center for the Study of Organizational Performance and Human Effectiveness, 1972. (ERIC Document Reproduction Service No. ED 068 515)

Tittle, C. K. Women and educational testing. *Phi Delta Kappan*, 1973, *55*(2), 118-119.

Tittle, C. K. Sex bias in educational measurement: Fact or fiction. *Measurement and evaluation in Guidance*, 1974, *6*(4), 219-225.

Tittle, C. K. Fairness in educational achievement testing. *Education and Urban Society*, 1975, *8*(1) 86-103.

Toior, A., & Brannigan, G. C. Sex differences reappraised: A rebuttal. *Journal of Genetic Psychology*, 1975, *127*(2), 319-321.

Vandenberg, S. G. What do we know today about the inheritance of intelligence and how do we know it? In R. Cancro (Ed.), *Intelligence: Genetic and environmental contributions*. New York: Grune & Stratton, 1971.

Vasquez, J. Measurement of intelligence and language differences. *Aztlan*, 1972, *3*(1), 155-163.

Vernon, P. E. *Intelligence and cultural environment.* London: Methuen, 1969.

Washington, W. E., et al. Petitioners vs. Alfred E. Davis, et al., respondents. Supreme Court of the U. S., No. 74-1492. (ERIC Document Reproduction Service No. ED 129 898)

Weber, G. *Uses and abuses of standardized testing in the School* (Occational Paper No. 22). Washington, D.C.: Council for Basic Education, 1974. (ERIC Document Reproduction Service No. ED 094 098)

Weckstein, P. Legal challenges to educational testing practices. *Inequality in Education*, 1973, *15*, 92-101.

Weiss, D. J. (Ed.). *Applications of computerized adaptive testing. Proceedings of a Symposium Presented at the Annual Convention of the Military Testing Association* (Research Report No. 77-1). Minneapolis, Minn.: University of Minnesota, Department of Psychology, March 1977. (ERIC Document Reproduction Service No. ED 137 360)

Weshner, M. C. Segregation — Alias: "special education." Paper presented at doctoral seminar, University of Georgia, Fall 1971.

Wick, J. W., & Beggs, D. L. *Evaluation for decision-making in the schools.* New York: Houghton-Mifflin, 1971.

Williams, R. L. The problem of the match and mis-match in testing black children, 1970. (ERIC Document Reproduction Service No. ED 073 131)

Williams, R. L. From dehumanization to Black intellectual genocide: A rejoinder. In G. J. Williams & S. Gordon (Eds.), *Clinical child psychology: Current practices and future perspectives.* New York: Behavioral Publications, 1974.

Williams, R. L. The Bitch-100: A culture-specific test. *Journal of Afro-American Issues*, 1975, *3*(1), 103-116.

Ysseldyke, J. E. Diagnostic-prescriptive teaching: The search for aptitude-treatment interactions. In L. Mann & D. Sabatino (Eds.), *The first review of special education.* Philadelphia: JSE Press, 1973.

Ysseldyke, J. E. Assessing the learning disabled youngster: The state of the art. University of Minnesota Institute for Research on Learning Disabilities. Research Report No. 1, 1977.

Ysseldyke, J. E. Issues in psychoeducational assessment. In G. D. Phye, & D. Reschly (Eds.), *School psychology: Methods and roles*. New York: Academic Press, in press.

Ysseldyke, J. E. & Foster, G. G. Bias in teachers' observations of emotionally disturbed and learning disabled children. *Exceptional Children*, in press.

Ysseldyke, J. E., & Salvia, J. A. Diagnostic-prescriptive teaching: Two models. *Exceptional Children*, 1974, *41*, 181-186.

Zirkel, P. A. Spanish-speaking students and standardized tests. *Urban Review*, June 1972, Nov. 1972.

SECTION IV

Tests and Decisions for the Handicapped

Ellis B. Page

PAGE, ELLIS B. Dr. Page received his doctorate from UCLA in Educational Psychology. For the past 15 years he has been Professor of Educational Psychology at the University of Connecticut. With special emphasis in research methodology, he has done post-doctoral work at Michigan, and at the M.I.T. Computation Center. Dr. Page was also visiting professor at Wisconsin, Stanford, Harvard, and Javeriana (Bogota). He has been elected to Fellow status in the American Psychological Association (Divisions of Measurement and Evaluation, and of Educational Psychology). His affiliations include the American Statistical Association and formerly President of the Division of Educational Psychology. Dr. Page has been Editor or Advisory Editor to five journals and authored numerious articles, chapters, and other papers, many dealing with the recent emphasis on the scientific establishment of policy.

INTRODUCTION AND READER'S GUIDE

The Law and The Guidelines

Educational law is one thing; educational action is quite another. Between the two events, the passing of the law and the behavior of the school, must occur a chain of intermediate events: the interpretation of the law in terms of practice; the study of the feasibility of the interpretation; the successive adjustments, reorganizations, retrainings, and redesign of administrative procedures; the self-monitoring and reporting — the reality testing. And early in this process (if not before the Law is written), there should be careful consideration of the provisions in terms of what is currently known. This present study aims to be such a consideration, from a limited point of view, of an important new Federal law.

The law studied is the sweeping "Handicapped Law", P. L. 94-142, and the part of particular concern for this study is Sec. 615-5c, which mandates that the participating States and their local agencies will develop:

> Procedures to assure that testing and evaluation materials and procedures utilized for the purposes of evaluation and placement of handicapped children will be selected and administered so as not to be racially or culturally discriminatory. Such materials or procedures shall be provided and administered in the child's native language or mode of communication, unless it clearly is not feasible to do so, and no single procedure shall be the sole criterion for determining an appropriate educational program for a child.

Obviously, this part of the Law is too cryptic to be immediately carried out in the local education agencies (LEAs). And therefore, after intensive study between 1975 and 1977, the Office of Education of the U. S. Department of Health, Education, and Welfare, issued a set of guidelines, published in the *Federal Register* of August 23, 1977, and titled "Education of Handicapped Children: Implementation of Part B of the Education of the Handicapped Act." The particular section of this "Implementation" dealing with the above quote are Secs. 121a530 to 121a534, generally headed "Protection in Evaluation Procedures" (and which shall be called PEP in this paper). The relevant parts of this Implementation will be briefly summarized:

Sec. *530* reiterates that testing and evaluation materials and procedures used for evaluating and placing handicapped children "must be selected and administered so as not to be racially or culturally discriminatory."

Sec. *531*: There must be a "preplacement" evaluation of each child.

Sec. *532*: Materials must be provided in a child's "native language" or "other mode of communication" wherever possible; be "validated for the specific

197

purpose"; administered by trained personnel in standard ways; directed at specific areas of "educational need" and not simply at general intelligence; "accurately reflect the child's aptitude or achievement" or other targeted factors, "rather than reflecting" the child's impairment (except where those skills are themselves the target); never consist of a "single procedure" as the "sole criterion". The evaluation should also be made by a "multidisciplinary team" including a specialist in the "suspected disability." And the child must be assessed in "all areas related to the suspected disability, including, where appropriate, health, vision, hearing, social and emotional status, general intelligence, academic performance, communicative status, and motor abilities."

Sec. *533* deals with placement procedures, and mandates that the LEA shall carefully consider many sources of information, including "aptitude and achievement tests, teacher recommendations, physical condition, social or cultural background, and adaptive behavior." The teams must include persons "knowledgeable about the child," the data, and the options. And the decisions must conform with the "least restrictive environment rules." Any placement decision must involve "individualized education program" (IEP).

Finally, Sec. *534* provides for "reevaluations" at least every three years, but more frequently "if conditions warrant" or if a child's teacher or parent requests it.

Structure of This Report

The present work is organized into four chapters to respond to these guidelines above. Since "fairness" in a system exists far more in the decisions made than in anything else, Chapter I examines what is known scientifically of decision processes in general, outlines the characteristics of formal decision analysis, and sketches out what is apparently needed in order to make such analysis function under a Handicapped Law. Particular attention is given to the central role of values in such decisions, and some ways are suggested for determining these values for such use. Here, as in Chapters II and III, recommendations from these analyses are largely saved until later.

Chapter II considers the difficulty of finding reliable and valid assessment methods for the Handicapped. It shows the effects resulting from choosing any extreme cases by some quota, either of observed score or true score, and gives particular attention to the vexed problem of reliability for difference scores, since identification of learning disability involves subtracting ability from achievement. And the question of reporting true scores for these differences is also explored psychometrically. When the reliability of ability is quite different from that of the specific achievement, an LD "decrement" may actually turn into an advantage! Some suggestions are made here for improving the

reliabilities, but most of the resulting recommendations are saved, once again, for Chapter IV. The question of appeals is also considered in Chapter II.

Chapter III touches on a number of technical matters related to the Law, especially those concerned with "racial or cultural bias." Often interpretation of "bias" seems most to depend on 1) sources of group difference, and 2) likelihood of remediation through differential treatment. Therefore some technical consideration is given to the questions of measuring heritability, the possible use of different ethnic norms, and possible procedures for guiding the foreign-language student, with the aim of acculturating that minority student into the majority. In Chapter III also is considered the question of accessibility of public records, and the consequent problems of test security.

Chapter IV is designed to put in one place all the major recommendations which seem to follow from the preceding chapters. The serious student of these questions, given both the time and the technical knowledge, may wish to work through the report, as the writer did, from the psychometric examination of the issues through the summary recommendations. A reader interested only in certain features may wish to consult the index of the report, studying those parts in particular, and then the recommendations at the end. For such a reader, the index will probably serve as an adequate guide. But for the reader who is mainly interested in tallying opinions on these questions as guides to conduct, Chapter IV will serve by itself. Since opinions will differ, however, from one investigator to another, it seems very important to gather, from Chapters I to III, some feeling for the sort of evidence and reasoning used in arriving at the recommendations.

Many have helped in studying these questions. Dr. James Ysseldyke kindly provided papers he had written on related concerns of the Law. My colleagues at the University of Connecticut, especially Drs. Isabelle Y. Liberman, A. J. Pappanikou, and John F. Cawley, have given valuable background in Special Education. And professionals engaged in the schools, Mrs. Linda H. Paananen and Mr. Joseph F. Stano, have helpfully shared their own experiences and investigations. Especially, Dr. Linda G. Mora, Education Program Specialist with the Bureau of Education for the Handicapped, has, together with her coworkers, provided valuable insights and suggestions for improvement of the earlier thinking. Nevertheless, the opinions and conclusions expressed in this report, together with any errors still present in fact or judgment, are entirely the responsibility of the writer.

CHAPTER I:
A DECISION SYSTEM FOR THE HANDICAPPED

The implementation of Public Law 94-142 requires that many decisions be made relating to handicapped children by each participating LEA. These decisions are large and small, affecting the way in which the entire system is to be administered, or the way an individual child will be diagnosed and placed. Examples of such decisions might be:

1) The LEA will decide about the structure of a particular *program* for classifying and placing and treating youngsters;

2) The LEA will decide about the particular *eligibility rules* for a particular treatment program;

3) The LEA will decide whether a particular *child* should be assigned in a particular way.

This paper will make a number of recommendations about how such decisions should be analyzed, and these recommendations will be in part based upon psychometric and educational beliefs; but also, to a substantial degree, upon the formal and highly developed theories of decision analysis, as used in the fields of management science and operations research. Unfortunately, however much Special Educators have written of "decision making," this theory remains little understood. If the reader is already conversant with formal models for decisions, then he may skip ahead to Figure 3. But for most readers, I urge attention to the introductory material. After illustrating the use of such decision models, this chapter will consider the central variables required for making such models function in practice.

Discrimination and Educational Decisions

It is frequently pointed out that "fairness" does not exixt in the tests themselves, but in how they are used; that is, in the decisions made on the basis of the testing. And when we consider applications of P.L. 94-142, we are concerned that there be evaluations of the identification programs and of the treatment interventions; and "evaluation" implies, once again, the relevance of the task to some subsequent decisions. Indeed, in evaluations texts, we find frequent reference to "decisions," yet seldom any attention to what is known about scientific decision making (Page, 1975).

Yet there exists a large body of work in the well-established discipline of operations research (e.g., Trueman, 1974; Wagner, 1968), and there is particularly useful and readily grasped structure in the sub-field of decision

analysis (Raiffa, 1968). The charge from Congress and from the Department of Health, Education and Welfare especially requires the conceptualization of a working system in which a myriad of decisions are made by LEAs and SEAs affecting the educational lives of the nation's handicapped. It is fitting, therefore, that we explore the points of contact between this comprehensive goal of our government, and the impressive techniques of such decision analysis. First, we shall design an abstract decision system for the implementation of the Handicapped Law. Then, we shall consider what sorts of information need to be generated in order to operate the system, with suitable protection against bias.

Decision Making For The Handicapped

It is intuitively recognized by most people that rational decisions depend on estimating certain variables: probabilities of various outcomes from the decisions; the likely benefits; the likely costs. Note that costs might be measured one way (such as time spent), and the benefits quite another way (such as pleasure expected). Yet it is clear that in practice, as individuals, we have little trouble handling these two kinds of value – or we could never decide whether to pay for a movie! When we come to these problems as professionals, however, we find only a trickle of research attempting to reconcile such different scales. To motivate this discussion, let us first assume that the values problem is tractable, and look ahead to the advantages of a formal system.

Decision analysis may be thought of as a sophisticated elaboration of that intuitive idea about probabilities, costs, and benefits. Any decision situation may be thought of as a "tree", with the top node referring to some immediate question, and the nodes below it referring to subsequent questions. These nodes are of two kinds:

- □ – square nodes, denoting *decision*, from which the descending branches are alternative *choices*, mutually exclusive; and
- ○ – circle nodes, denoting *probability*, from which the descending branches are alternative *events*, mutually exclusive.

These nodes may multiply into large structures, but all must conform to the definition of an upside-down "tree," having a common source at the top, and complete separation of all branches. Each of these nodes is calculated separately, beginning at the bottom of the tree, and working up till reaching the top-most node. If the structure of the tree is a good match to the real world, and if the numerical values are estimated well for each part of the tree, then the decisions are in fact automatic, and can be made efficiently by a computer. Error-free tree design, then, leads directly to error-free decisions. And this statement remains true even when there are large doubts about the likelihood of future events – so

201

long as the probability estimates, themselves, may be considered accurate. Still a further claim can be made for the tree. Given the same human estimates of the parameters in the tree (the structure, probabilities, costs and benefits), the algorithm will always match or better the human decision maker.

Figure 1 shows the simplest kind of decision node, and illustrates how the optimum is decided. We shall consider, here and in subsequent problems, that we are choosing among three programs or "Plans" for the administration of the Handicapped system. We have estimated "Utilities" for the three Plans, by methods we shall see. The rule at any decision node is very simple: We select the Plan which yields the highest estimated Utility. Of the three listed U values (14, 20, 14), 20 is obviously the maximum. Therefore $max(U_i) = U_2 = 20$. The value of the decision node becomes 20, and the two rejected branches, Plans 1 and 2, are "folded back", a process indicated by two barrier lines athwart each branch.

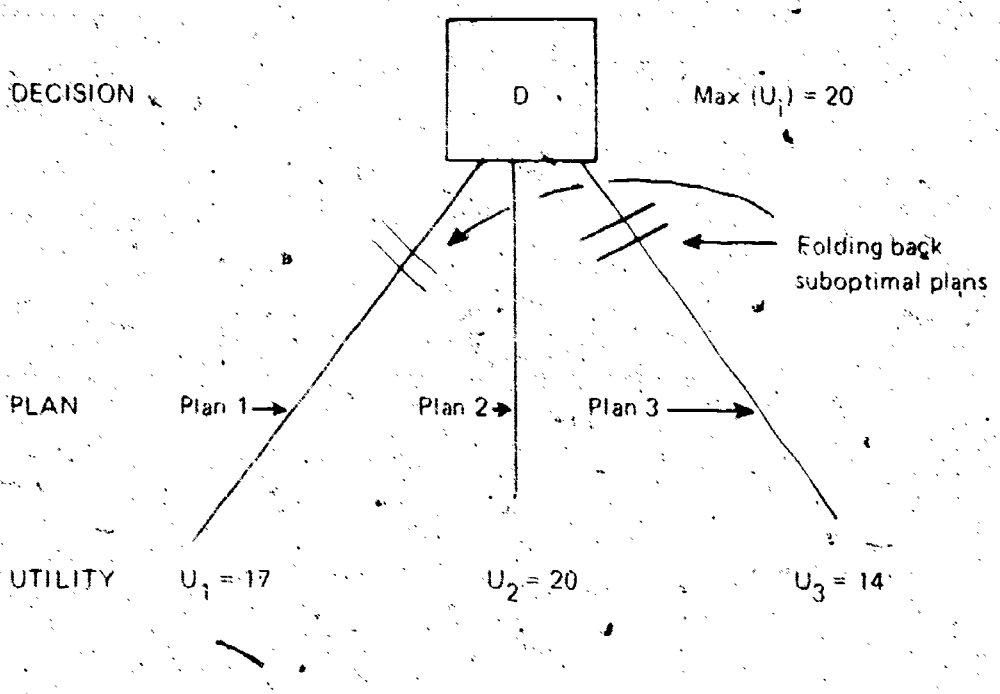
As we have noted, the other major type of node is for probabilities, and we show a probability node in Figure 2. In this illustration, we are using the probabilities to estimate the costs of measurement for one of the Plans. There are just four classifications of pupils here (for purposes of this Plan). Each classification c has its own probability p_c and its own cost of measurement M_c. We find the value of this probability node by "averaging out" the descending branches: we multiply each M_c by its probability and sum across the branches of that node. Thus the value of the node in Figure 2 is shown as 3.30 of some appropriate measure.

What is the meaning of "probability" as used here? For decisions in general, it may be of either classic kind, either based on subjective prediction (as for unique future events) or based on long-range frequencies (as for the result of past experience, or even of some statutory or other quota). The algorithm is indifferent to the source, but obviously the "probability" values will be much more precise if the system mandates the distribution, as in filling classes of predetermined size.

For a more complete tree of decision for the handicapped, what elements will be needed? We are interested in *decisions*, made on the basis of *evaluations*. And between these two ends of the chain we need *plans* (which will be object of final selection), *classification* (applied to the prospective pupils), *treatments* (appropriate to the classification), and *outcomes* (within treatment and classification). Such a chain is shown in Figure 3.

In this illustration, there are assumed to be three plans. The second, Plan 2, has a system of four classifications of pupils. For the third, Classification 3, there are three relevant treatments. And for the second, Treatment 2, there are three outcomes. These three outcomes carry the evaluations, E_1, E_2, and E_3. Note that, by making certain other assumptions, two levels of the tree could be eliminated: We could collapse levels II and III, if there were only one authorized treatment

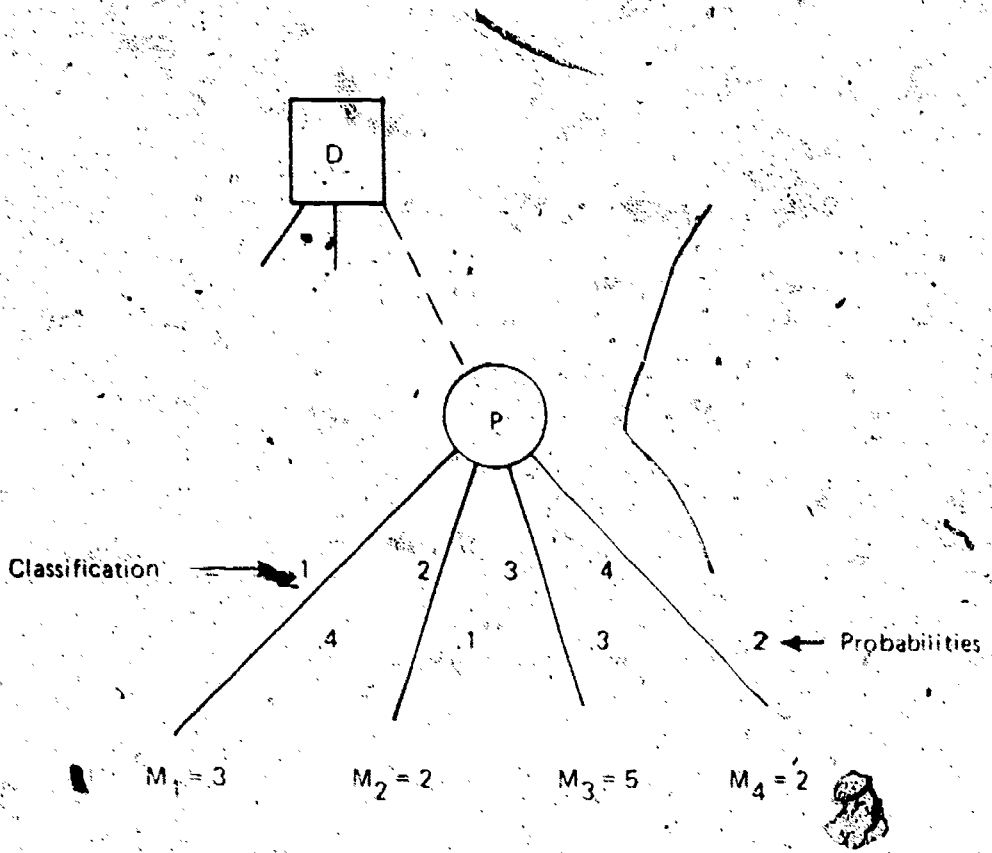
FIGURE 1
CHOOSING A PLAN



At the highest level, a rational decision consists of selecting that plan which maximizes the estimated Utility. Those plans which are suboptimal are "folded back," an action denoted by barrier lines. The "value" of the decision itself then becomes the utility of the plan selected, in this case $max(U_i) = U_2 = 20$.

FIGURE 2

CALCULATING COSTS OF MEASUREMENT

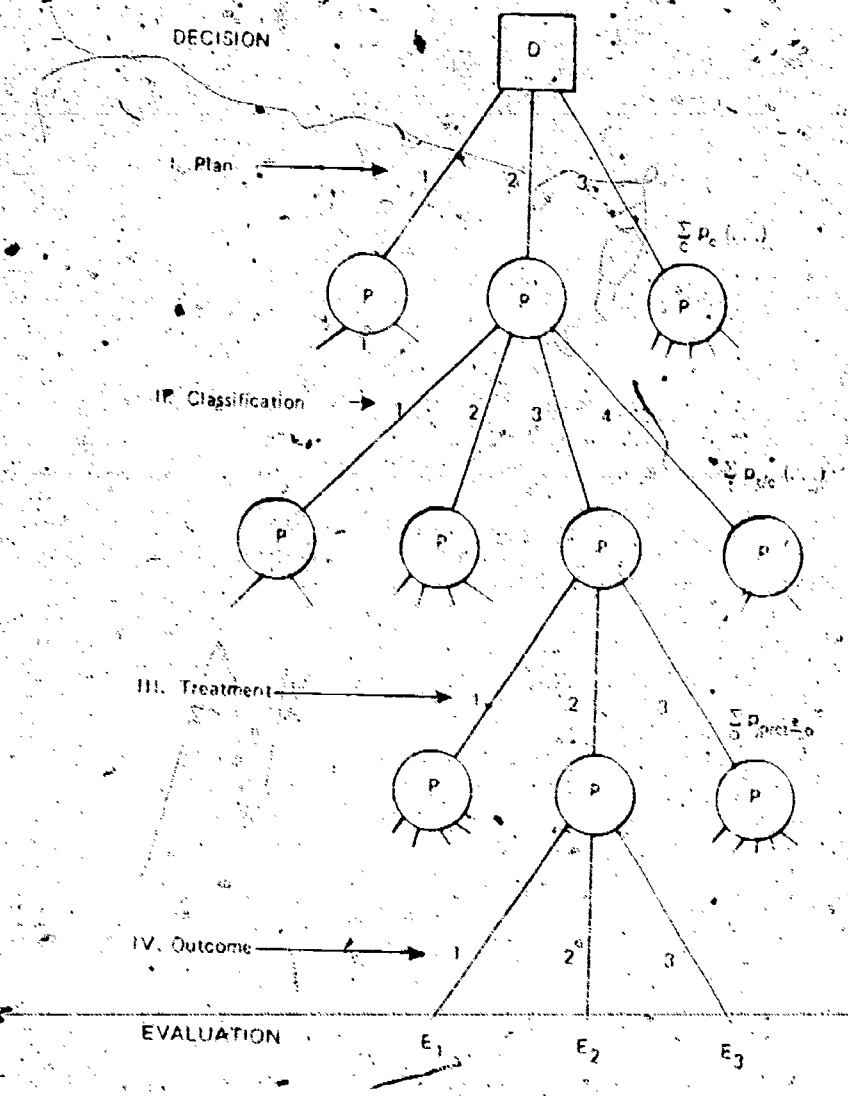


$$\sum_c p_c M_c = (.4)3 + (.1)2 + (.3)5 + (.2)2 = 3.30$$

In decision analysis, each probability node may be calculated, whether for benefit or for cost, by multiplying the probability of each alternative (in this case pupil classification) by the terminal benefit or cost (in this case the cost of all measurement for that classification). Calculations are shown for token estimates.

FIGURE 3

DECISION TREE FOR HANDICAPPED PROGRAMMING



The decision is which of three programs or plans to adopt on a system-wide basis. To implement the tree, the system needs estimates of the evaluations, and the probabilities of three levels in the tree: pupil classifications, treatments for each classification, and outcomes for each treatment. (See the following pages for illustrations.)

PROGRAMMATIC ILLUSTRATION OF DECISION TREE

Figure 3 will be perhaps better understood if fleshed out with illustrative branches for decisions and events. Therefore let us say that the levels of the tree are as follows:

DECISION. There are just three plans being considered by the LEA for administering the Handicapped Law. These plans, let us say, have some major differences in structure (which will not be detailed here), and the LEA therefore wishes to evaluate their potential operation, and choose one of them.

LEVEL I. PLAN. We shall simply call these Plans 1, 2*, and 3, to avoid a cumbersome sketch of each.

LEVEL II. CLASSIFICATION. Here we assume four types of general identified students. For simplicity, these are:

1. physically impaired;
2. emotionally disturbed;
*3. learning disabled;
4. culturally disabled.

LEVEL III. TREATMENT. Here we postulate three treatments as alternatives for any LD child:

1. regular classroom;
*2. self-contained special education classroom;
3. regular classroom with resource room placement.

LEVEL IV. OUTCOME. Under Treatment 2, we here assume three possible outcomes (as measured 1 year later):

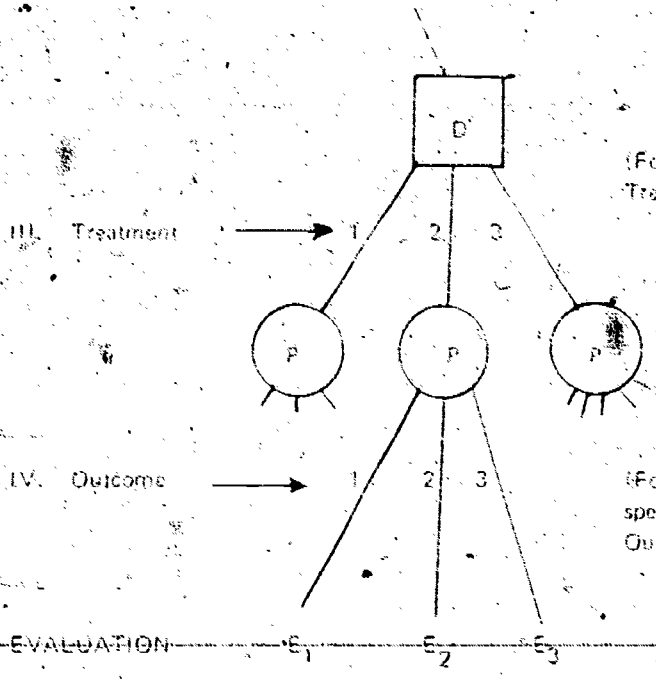
1. move to regular classroom;
2. remain in self-contained special education class;
3. move to resource-room treatment.

EVALUATION. The decision tree requires this most important level of evaluation of the three outcomes possible within this treatment, classification, and plan. As here presented, these could depend on assessment of pupil condition implied by the outcomes, or could depend on weighted sums of test scores or other measures (as in the "bentee" strategy to be described).

**Note:* Asterisk marks the particular path through the tree illustrated in the Figure 3.

INDIVIDUAL
ILLUSTRATION OF DECISION TREE

Figure 3 is illustrated (p.) for a selection of programs. Within that decision, the pupil classifications are treated as probabilistic events, based on either past or predicted long-run averages of such classifications, treatments, and outcomes. For the *individual pupil*, however, the nodes of this tree may often change from probabilistic to decision nodes. For example, once a child is classified as "learning disabled" (see *LEVEL II. CLASSIFICATION* of the preceding page), then there may be a sub-tree of the following sort:



In this illustration, the probabilities of outcome, within each treatment choice, will now be different for the individual from what they were for the group as a whole. These will again depend on past experience and future prediction, but now based upon this student's *individual* profile.

for each classification. And for some, the "outcome" would not simply be a state description, but would itself be the numerical evaluation. But the tree of Figure 3 may be better, since it makes these actions more explicit.

Many trees in decision analysis have a number of different decision nodes, but here there is only the one. All other nodes are reflecting different events which could occur, or which would be ordained to occur by quotas in the system. Thus, each node below the top would be averaged out, beginning at the bottom of the tree, and working upward until every node had its own value except the top, decision node, and that one would be solved by selecting the Plan with the largest benefit, all things considered. The algebra of value calculation for each node is suggested by the summation operators to the right of the Figure.

Calculating the Utility of Plans

If we wish to understand the structure of the system from a different perspective, we turn to Figure 4. Here we have said that there are just three classifications, five treatments, and four outcomes, so that our $c \times t \times o$ matrix has $3 \times 5 \times 4 = 60$ cells. Some of these would by empty: If a pupil is diagnosed (classified) as gifted, he is unlikely to be prescribed (treated) as Learning Disabled. Furthermore, the "outcomes", as we have suggested, might be transformed into some appropriate continuous variable (more of this later), which would turn the matrix into just two dimensions (classifications by treatments) within any single Plan. But in this illustration, there is an evaluation, E_{cto}, for each meaningful cell. There will be more about such evaluation, too.

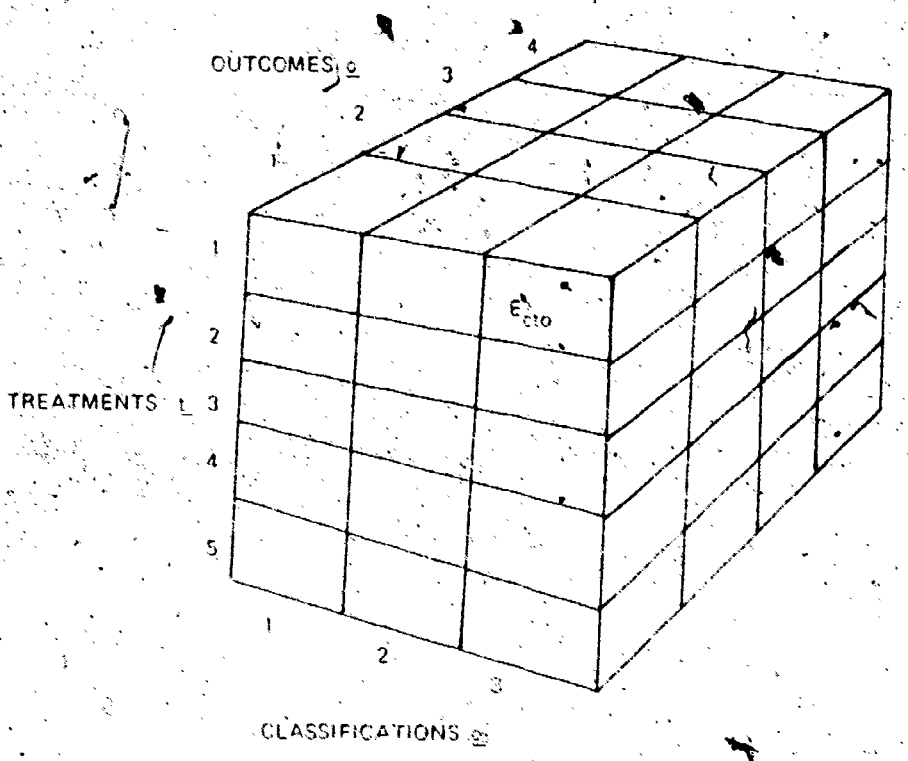
But we are going to select a decision in terms of its estimated "Utility" – a general term from other disciplines meaning the attractiveness of a choice; the units are not constant across studies, but may be designed for a situation, as we shall see. For the present, let us assume that we have such a measure. Then a general formulation could be that Utility is equal to the benefits, less the costs.

For a testing program, there is an excellent, seminal work by Cronbach and Gleser (1965), which considers a program much like that of Figure 3, in which the principal cost to be considered was the cost of *measurement* (p.24). For such a program, Utility would be equal to the overall evaluation of the program less the cost of measurement within the program. Formally, we can set forth the algebra as in Figure 5.

Granted, this level of expression may seem overwhelming to many practitioners in LEAs and SEAs, and it is not suggested that they work directly with such formulations. But creating complex trees is easy computationally (e.g., Findler, Pfaltz, & Bernstein, 1972), and computer aids to decision making could (and

208

FIGURE 4
THE EVALUATION MATRIX



For a given plan, it is possible to construct a three-dimensional matrix according to pupil classification, the treatment alternatives for a given classification, and the predicted outcomes for each such classification/treatment combination. For each feasible cell, one evaluates the outcome considering both benefits and program costs.

209

FIGURE 5
CALCULATING THE UTILITY FOR A PLAN

CALCULATING UTILITY FOR A GIVEN PLAN

The overall "utility" of plan may be given by the formula:

$$U = N \sum_c p_c \sum_t p_{t/c} \sum_o p_{o/ct} E_o - N \sum_c p_c M_c$$

where p_i = the probability of the ith event,

U = utility of the plan

N = number of pupils for whom plan is designed,

c = the pupil classification from measurement,

t = the treatment selected,

o = the outcome from the treatment,

E_o = the evaluation of the oth outcome,

M_c = the cost of measurement for the cth classification.

The value of U is dependent on the benefits, less the costs, and therefore assumes construction of a single scale of measurement.

possibly should) become as commonplace as student scheduling programs, once their uses are sufficiently appreciated.

But the most serious problem with implementing such algorithms in education is probably in the assignment of reasonable values. And it is to this problem that we turn our attention.

General Values in Education

Uses of operations research have thrived most in fields where there was a common, agreed-upon measure of utility, such as dollar profits, or distance and time costs. In education, we are often thinking in terms of *educational* benefit, with no unit or scale readily available. Consequently, large major types of problems are simply without any investigative approach as commonly understood. Some works have treated decisions within education, but have sidestepped the fundamental questions of value (VanDusseldorp, Richardson, & Foley, 1971). Others have deeply analyzed the theoretical considerations underlying multidimensional and curvilinear value systems in decision making, but have few applied suggestions for educational practice (Keeney & Raiffa, 1976; Wilcox, 1972).

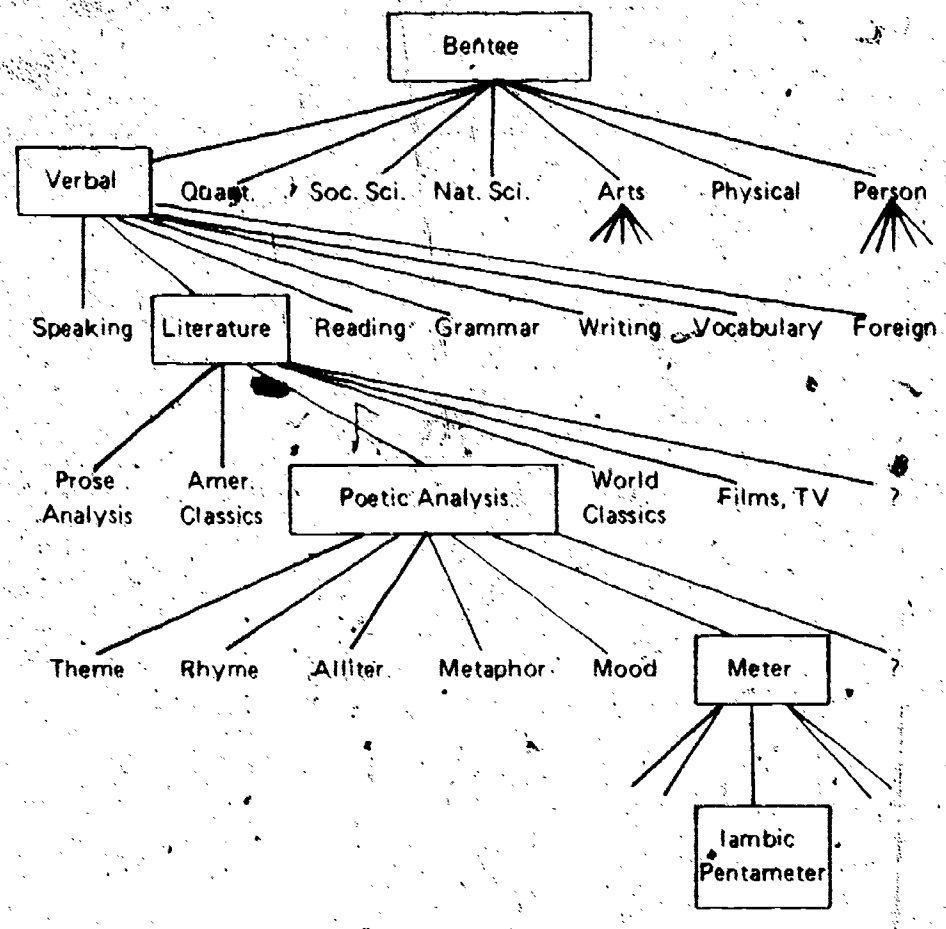
Part of our weakness in understanding complex values is the obverse side of one of our strengths: our sophistication in multivariate analysis, where we discover appropriate weights for optimizing a prediction. The problem in evaluation begins at the other end: rather than *discovering* our value-weights, we *invent* them. The ultimate values are dependent on some quite subjective process; the advance must come in the manner of collecting them, combining them, and using them in decision situations. Yet a good review of techniques for multiple measurements (e.g., Cooley, 1971) will take no notice of the apriori values needed for decisions.

To meet this problem, we have attempted to create a scale of overall educational benefit. We have begun by defining a unit of measure of such benefit, expressed as a T-score of educational accomplishment, scaled in the usual way with mean of 50 and standard deviation of 10. Scores, then, range from about 20 to about 80. And this "benefit T-score," or *bentee*, is a function of a weighted sum of the values of certain variables multiplied by their measurements for an individual or group. For example, if the total value of the bentee for a high-school senior is taken to consist of some function of verbal, quantitative, social studies, natural science, arts, physical, and personality, we may display those as branches from the bentee node, and each may be in turn broken down. Such a tree is shown in Figure 6.

This tree has some remarkable features (as described by Page, 1972b; Page, 1974b; Page & Breen, 1976). It borrows many of the mathematical qualities of

211

FIGURE 6
TREE OF EDUCATIONAL VALUES



Source: Page (1974b).

As analysis moves from the general to the specific, a shift is made from societal to expert opinion, and from value-space to test-space.

the probability trees we have been examining. The values for each set of branches may be made to sum to 1.00 for any node. Then these may be multiplied down the lines to calculate the bentee for a deeper node. Values where they are repeated may also be summed for a total appraisal of the value of a given knowledge or skill.

The most striking quality of the bentee tree, however, is its philosophical completeness. Within only a few generations (seven in this illustration), one may move from the top-level philosophical values for society, the system, or the individual, down through successive divisions to the lowest, most direct test items or specific behavioral objectives of instruction. It provides, then, a way to subsume much of education within a single system of value.

How to perform these evaluations, then, becomes of concern, and the answer is fairly straightforward: In effect, we have appropriate judges vote on them. Various methods have been investigated by researchers, and we have given our attention especially to two, and one of these survived as more convenient and less expensive. It is called the *token* method, and works by asking the judge to "spend" 100 tokens among a number of alternatives, according to the way he or she (the judge) feels is a proper apportionment of value.

An empirical investigation of the values so assigned by 101 judges (half professional educators and half laymen) showed that, in general, such allocations of value are made quickly, without apparent strain. Furthermore, despite wide individual differences, there was a clear agreement between the educators and the laymen about the top-most values (Page & Breen, 1974). These findings were partially confirmed in an applied setting with the U.S. Navy (Page, 1976b), but certain artifacts of judge behavior were also noted. In another, more theoretical article, it was shown how such tokening could be used in the design of a curriculum (Page, Jarjoura, & Konopka, 1976). Of course, as subjective judgments, such weightings are vulnerable to the same sorts of variables which affect ratings, but they are no more vulnerable than such other subjective processes, and they have the great virtue of being out in the open, where they may be appraised for reliability and representativeness of some target population of judges. In addition, use of the bentee strategy permits the incorporation of the values into decision-making algorithms which are, under certain assumptions, error free. In short, such a strategy *employs* subjective judgment, but *tames* it and limits it to those functions where it is irreplaceable by objective techniques.

Values for the Handicapped

Let us consider some cases of possible decisions affecting the planning for handicapped:

Case 1. Despite the "least restrictive environment" phrase, the absorption of the physically handicapped into regular classrooms will clearly have liabilities as well as advantages. Depending on the nature and severity of the handicap, such integration may disproportionately spend the educational resources of the classroom (such as teacher time), causing a certain loss in cognitive learning to the normal majority. How may such philosophical and ideological questions be resolved? How may the decisions be arrived at?

Case 2. Just a certain percent of the student population may be supported by the Federal Government within the provisions of P.L. 94-142. It is unreasonable to believe there will be no limits to such support, from any source. But there are many more who, depending on the personnel doing the evaluation, *could* be claimed in need of such remediation. They vary widely in type and degree of deficit: physical, sensory, and the entire range of Learning Disabilities as defined by the Law. There are, of course, various special interest groups, such as teachers, parents, and specialized personnel, all of whom have their particular concerns and who bring as much pressure as possible on the decision makers. Is there any more professional and more promising method of allocating the resources?

Case 3. A number of students, of apparent normal aptitude, are below standard in both English and math. A proposed math program would, from best estimates, raise math performance by four points (in T-scores) over the present program. But the time would be taken from English, which would apparently lose two points. Is the combination gain-and-loss desirable or undesirable?

The word "crisis" is overused, but in each of the three cases it seems reasonable to say that there is a crisis — or at least dilemma — about the legitimacy of any procedure ordinarily employed. Administrators will come to some decision, often on the basis of whichever pressures seem strongest at the decision moment; and when such pressures are much in conflict, the administrator is forced to make decisions unpopular with many, and virtually impossible to rationalize. Note that in each case there is an underlying question of *fairness* to the pupils concerned; therefore, any use of tests in arriving at such decisions must be evaluated in terms of protection against discriminatory misapplication.

Putting the handicap in perspective.

The first problem is to put all of the pupil profile, not simply his trait of deficit, in the decision framework. Otherwise, it will be very difficult to make any judgments concerning limited resources, and competitive handicaps. Thus, on

214

the basis of values established locally or taken from research already done, bentees (or some other scale for combining scores) are calculated for each student. Let us call this combined score B_i for the ith student.

Plotting the Production Function.

What we are interested in is now clear, in a general way: it is improvement in B_i for our handicapped students. Before we can make decisions about our system, we need some estimates of the probable success of our methods. For our purposes, we especially need to know how much improvement is expectable in B_i depending on the level of effort expended. Different handicaps will of course require different sorts of treatments, so the methods themselves will not be comparable. But most of these treatments will spend a common and limited resource: the professional *time* concerned. For illustration, let us set aside other costs, and agree that we have a total time T available for remediation, and that we shall spend time for each pupil, t_i, such that

$$\sum_{1}^{N} t_i = T,$$

where N is the total number of handicapped students.

For each student, from the best information we have available, we may estimate the values of a production function, like that one shown in Figure 7.

This figure shows an assumed relation between the time spent on a student (in remedial work) and the progress made. It assumes that three individual parameters describe the curve. The formula is that of a fairly familiar growth curve (e.g., Atkinson, 1972; Page, 1973):

$$B_i(t) = \alpha_i - \beta_i e^{(-\gamma_i t)}$$

where

$B_i(t)$ = bentee score for student i at time t;

α_i = maximum possible bentee of student i, assuming all time T allotted to him,

$B_i(0) = \alpha_i - \beta_i$ = starting bentee of student i, before extra time is spent,

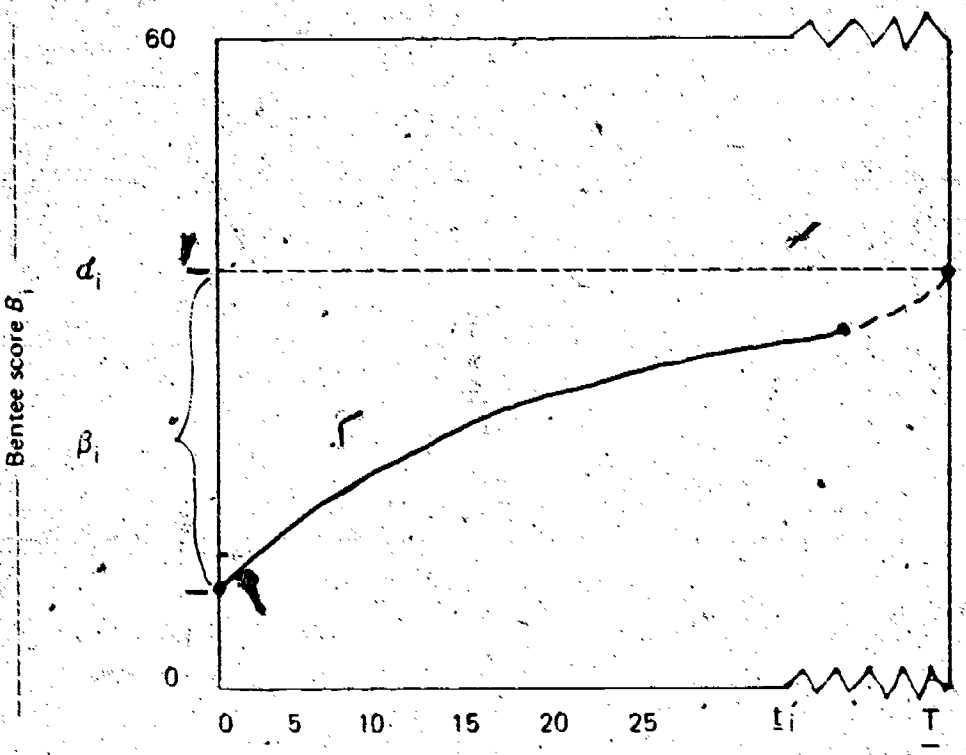
e = the base of the natural logarithm,

γ_i = growth function for student i.

At the time of making a decision about time allocation, we will have N students eligible for some help under the Law. For each we are able to estimate $B_i(0)$, the current bentee before beginning treatment, on the basis of testing. We must estimate α_i and β_i either from data on others with similar type and degree of handicap and their growth histories with treatment, or from the data collected on student i's improvement from the first hours of intervention (the procedure suggested by Atkinson for first-grade reading work for disadvantaged students).

FIGURE 7

PERFORMANCE AND RESOURCE ALLOCATION



Measured in overall terms, performance is described by the bentee B_i (t) for the time allocated to student i. Selection of time allocations (t_i^*) for each pupil would depend on such estimated growth functions and on the philosophy of the system.

Once having estimated the growth for each candidate, we now have the opportunity of deciding the allocations, according to the system's philosophy. It is now possible, at least, to state operationally what goal is being pursued. Let us look at some alternative goal statements, with an eye toward the "fairness" of the allocation:

Goal 1: Maximize $\sum_{1}^{N} B_i(t)$. This would have the effect of producing the highest mean performance for the population of interest. It would probably imply neglecting the lowest achievers, and it would surely mean neglecting those with the lowest growth parameters γ_i. Therefore, it is sometimes termed an "elitist" [redacted]

Goal 2: Minimize the variance of $B_i(t)$. This would tend to neglect the most rapid learners, those most apt to benefit from the system, even among the handicapped. It would be apt to neglect the slowest gainers, but would otherwise emphasize help for those with the lowest beginning scores. This is sometimes termed "egalitarian."

Goal 3: Maximize the inclusion of students with low input scores $B_i(0)$. This would ignore the individual parameters of possible growth, and would instead concentrate on those with poorest present status, regardless of probable gain from treatment. This choice is not often articulated, but seems to be close to the spirit of some citizens and lawmakers. This might be called a "compassionate" approach.

Goal 4: Minimize the variance to t_i. This would ignore all the parameters of growth, once the pupils were identified as handicapped, and would distribute the time equally across the pupils. If Goal 2 is egalitarian in *outcome*, then this is egalitarian in *allocation*.

Goal 5: Maximize $\sum_{1}^{N} B_i(t)$, under the constraint that each identified handicapped pupil will have some minimum time $k\frac{T}{N}$. If $\frac{T}{N}$ is the average time available for each handicapped pupil, then k is some fraction, such that every pupil will receive at least k of this average time. Beyond that minimum, the resources would be allotted to those who would most gain from the expenditure. Goal 5 would have the virtue of continuing to monitor and give some attention to each pupil, while concentrating on those who are believed to be the best candidates for remediation.

Undoubtedly there are other feasible goals as well. These are not given here with the intention of deciding which one is "best": that is not the role of a psychometrician. Rather, they are presented as an attempt to recognize that variables need to be taken into consideration when decisions are made on the basis of test information. In this analysis, the "fairness" of an individual decision would appear to depend on the "fairness" of the system. We should recognize

217

that resources for remedial education will not be without limit, regardless of the level of government. Thus, the problem is inevitably how to allocate those resources which are available. And it is difficult to imagine any reasonable system which does not recognize the informational needs which have been exhibited by this analysis.

The Complexity of the System

For those used to customary procedures, the demand for information may seem excessive, and the difficulty of the algorithms, compared with what we are used to, may seem severe. The response to these objections goes like this: Exactly which goals are we currently trying to optimize? And exactly what information do we need in order to do so? And just how will such information enter the system? This response emphasizes that *there is no alternative theory of rational decision making*. Anticipated benefits count on general predictions about present and future status of pupil performance. Utilities count on value weightings of anticipated benefits. Decisions count on some balancing of utilities against costs. Student selection counts on the principle of optimizing some function of the available variables. Lacking an alternative theory of decision making, we must recognize that, without the prior information available to us, we are currently making judgments which are unknowably sub-optimal. True, as Tillett (1975) found in a comparison of current vs. optimal teacher assignment, good professional decisions may not be very distant from optimal ones. Yet the absence of such information renders illusory any attempt at comprehensive evaluation.

Furthermore, there are reasons for optimism in such applications, if we look at the sort of complexity which is now handled in artificial intelligence systems (Minsky & Papert, 1969; Slagle, 1971), large data-retrieval sets (the National Longitudinal Study), and the massive mathematical programming systems which can optimize a thousand variables. Granted, more investment will be needed in such applications to educational decision making; but the additional sums needed for research and systems engineering, are very small compared with the expenditures for the application of P.L. 94-142, and even for the testing programs and committee operations necessary to guide those applications.

Above all, it is necessary to recognize that human cognitive systems, whether of the individual or committee decision maker, are simply unable to compete with effective algorithms. Hills (1971) made this point very well in a review of the literature on what is often called "statistical vs. clinical prediction." His conclusion on this point is worth noting at length:

> [I]f expert judges are to be used in the selection process, present data suggest that they not be allowed to make the final evaluations upon which decisions

> are made but that they be used as expert observers producing data that can be introduced into prediction equations or other statistical combination procedures that generate the final evaluations . . . Those applicants who exceed a cutoff point on the estimates of performance should be accepted with no further human intervention. (p. 697)

In other words, once the human beings have performed their essential human roles of providing values and judgments not otherwise obtainable, then the precision of the algorithm should be allowed to operate, as noted elsewhere (Page, 1974a).

In this chapter, I have attempted to apply some well-established theories of decisions, and some other research of probable relevance, to the problems of making decisions in the Handicapped Program, whether about the structure and administration of programs or about the placement of individual youngsters.

Note, the multivariable nature of these decision systems assures that the Law will be honored and that "no single procedure" will constitute the basis for inclusion or exclusion in the Handicapped system.

Relevant to this question of values is the operation of the teams of professionals called CETs or PETs. These teams may be considered analogous to juries, when one is considering only the classification decision; and this relation is explored in Chapter III. Other questions of decision making are dealt with in Chapter II, primarily devoted to questions of reliability.

Material in this chapter has had to be often quite abstract and general, and some of it too technical for most of the decision-makers in the schools. It would be very inappropriate, however, to ignore such material on that ground for, as Dewey put it, "A good theory is the most practical thing of all." And the theory of decision analysis, considered mathematically, is demonstrably sound; in fact, *there is no respectable alternative theory about decision-making*. Yet to take advantage of such reasoning, it is not necessary that everyone understand it, any more than one needs aeronautics to be a successful pilot. It *is* necessary that the more technical people confront such aids and study them hard for their applicability in education.

CHAPTER II:
PROBLEMS OF RELIABILITY FOR THE HANDICAPPED

This chapter is concerned with many of the questions bearing on the reliability of the assessment procedures for implementation of the Handicapped Law. It is intended to lay the technical background for recommendations by considering such reliability from the most classic perspectives, and then looking at the implications for the special sorts of measurement required of the Law.

219

First, the nature of "true scores" is examined for the situation of parallel forms, since these are analogous to the use of different but closely related instruments for measuring the same ability or achievement. Then we examine questions of selection, of extreme groups such as those eligible for assistance under the Law, given the discrepancy between observed and true scores. Then we look at the questions of LD and other assessment, where we judge "specific disability" by comparison with some "general ability" measure — questions loaded everywhere with statistical booby-traps. The reliability of difference scores is seen to depend on the respective reliabilities of ability and achievement scores, and tables are provided to summarize these relationships. (Some of these considerations may seem forbiddingly technical to the school professional, who is encouraged to skip ahead, yet they may well be worth retaining for eventual use in computerized systems of interpretation.) Finally, some recommendations are made for improving the use of tests, especially by enhancing the reliabilities and validities of the overall scores.

Reliability of Specific Tests

For a number of reasons, the reliability of tests is of particular importance in designing systems for the selection of the handicapped. Ysseldyke (1977) has recognized this importance, and has made a number of recommendations about the use of such tests. One is that we should perform most of our calculations not with the observed scores, but with the regressed "true scores," based on estimates of the reliability of the tests. Let us consider some of the implications of such a policy.

Let us imagine the cleanest data for estimating the reliability of a test score: the case where we have two forms of the test, assumed to have been randomly drawn from a universe of test items, such as pictured in Figure 8. When the forms are given on separate occasions to the same set of subjects, there will be a correlation between them, which we term r_{12}. Of course, this relationship will be affected by all of the many influences bearing on test performance (Stanley, 1971, p.364): lasting and general characteristics of the individual; lasting but specific characteristics of the individual (such as the handicap in question); temporary but general characteristics of the individual (such as fatigue); temporary but specific characteristics of the individual (such as comprehension of the specific test task); systematic or chance factors affecting the administration of the test or the appraisal of test performance (such as unreliability or bias in grading or rating); and otherwise unaccounted variance (such as luck in guessing answers). Notice that all of these tend to weaken r_{12} except the first

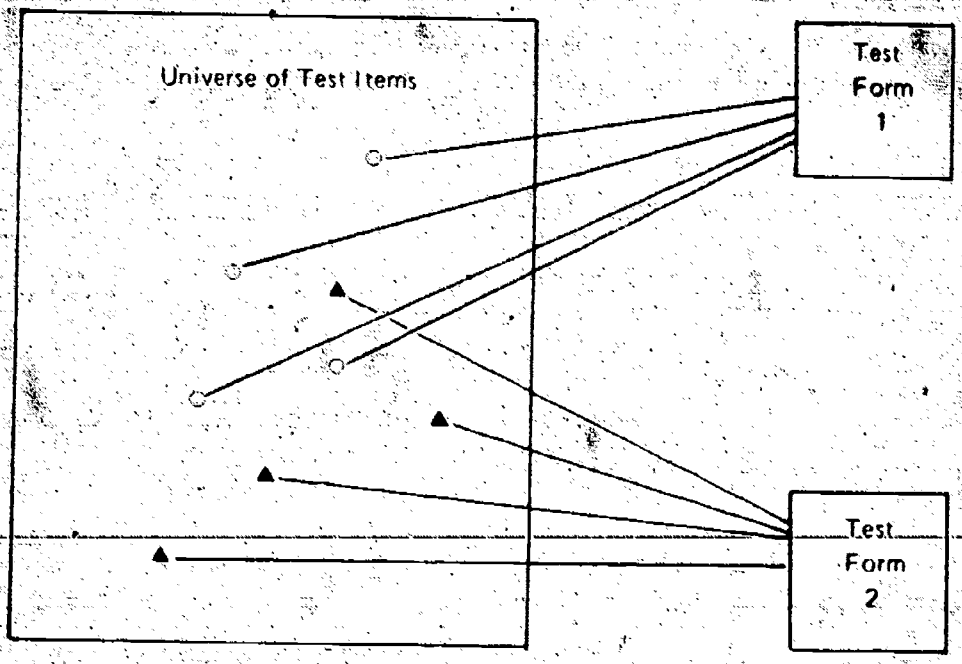
220

two: lasting characteristics, either general or specific, of the individual. And the other influences are generally those which, for purposes of most testing programs, contribute to the "error".

$$\sigma_e^2 = \sigma_x^2 (1 - \rho_{ff'}),$$

where $\rho_{ff'}$ is the population correlation between forms 1 and 2.

FIGURE 8



Each test form is assumed to be a random sample from a universe of test items concerning the ability or skill measured.

221

Now, in the above Figure, it is noted that there is no direct causal relation between the content of the two forms. Rather, the content is similar only because of the homogeneity of the universe of content from which they were drawn. This point is seen in the path diagram of Figure 9. The causal relations are denoted by the arrows drawn from the true score (that score achieved by the subject on the universe of all items for the trait) to each of the test forms. The curved line connecting the two test forms denotes a relationship which is only fortuitous, in being a reflection of their common source. In such a case, as the literature on path analysis makes clear (e.g., Blalock, 1971), we know that any observed relation between the forms must depend on the paths connecting each to the true score, and must be a product of those two. Thus,

$$c = ab,$$

and if a and b are assumed to be equal, then

$$a = b = \pm \sqrt{c}$$

The causal path from the true score to the score on either test form, then, may be estimated as $\sqrt{r_{12}}$. This has the result that, in standard form,

$$Z_3 = p_{31} Z_1 = \sqrt{r_{12}} Z_1$$

However, when we are working only with observed score values (such as are expressed in T-scores, grade equivalences, etc.), then we must remember that true scores will have a smaller variance than the observed scores, because of the removal of error. That is, according to one definition of reliability (Stanley, 1971, p. 374),

$$\rho_{tt'} = \frac{\sigma_T^2}{\sigma_x^2} \quad \text{hence} \quad \sigma_T = \sqrt{\rho_{tt'}}\,\sigma_x$$

that is, the population correlation between forms is the ratio of the true-score variance to the total observed variance. To convert the true score to units of the observed-score, then, we would say that, in the path diagram,

True score (in standard units of Test Form 1) = $r_{12} Z_1$, or, for the usual computational formula:

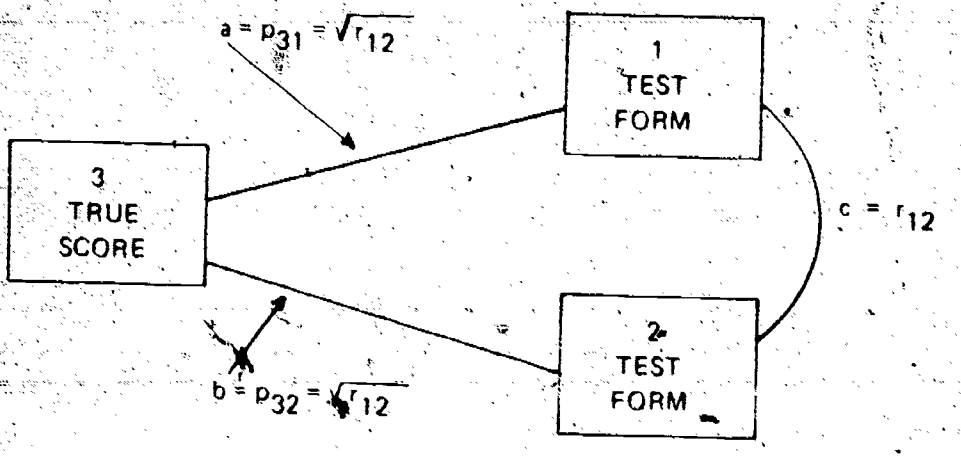
$$T_i = \bar{X} + r_{12}(X_i - \bar{X}),$$

where T_i is the estimated true score for the ith student and X_i is his observed score.

Now, it is clear that T is merely a linear (first-degree) transformation of X. And

FIGURE 9

PATH DIAGRAMS FOR RELATION BETWEEN TWO TEST FORMS AND A TRUE SCORE FOR THAT TEST



Any relation between the two observed test scores (such as r_{12}) is caused by their common relation to the true score, since they are considered samples of the universe as in Figure 8. Thus, the path of c is the product of the paths for a and b. Therefore the relation of each form to the true score is the square root of the observed inter-form agreement.

223

this recognition leads us to a highly relevant observation about selecting students for remediation on the basis of such true score, as recommended by Ysseldyke. Whether the recommendation will make any difference depends upon how we intend to determine who is included in remediation. The two cases are shown in Figure 10, and another Figure 11.

In much of the discussion which follows, we shall be assuming some cut-off mandated by the limited resources. The word "quota" is currently unfashionable, yet it must be recognized that, in a field as fuzzy as learning disability, there must be in fact quotas, whether explicit or disguised. Estimates for specific "reading disability," for example, range from 2% to 20% or higher. It is obvious that, in fact, hardly any student achieves his/her "real potential." To give statistical analysis some concreteness, then, I have often assumed a quota of just 2%. In practice, this will be larger or smaller, depending on the nature of handicap, the expense of the treatments being considered, and the budgets of the supporting agencies. In the following, therefore, the "2%" should be considered only illustrative.

In case (a), we assume a quota selection including just the bottom two percent of the population in our remediation. In such a case, transforming to true scores has no effect on selection, since under linear transformation there will be no change in rank order of any selected.

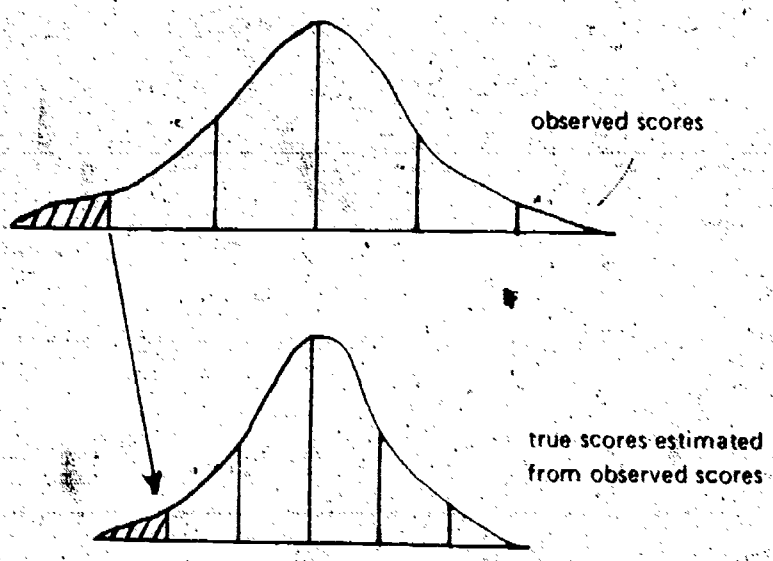
In case (b), we assume the only condition which would make true-score transformation meaningful: that it will change the composition of those selected for treatment. We observe that, with a cut-score in terms of the observed distribution, the *number* of those selected can be drastically altered, depending on the estimated reliability of the selection test. In the Figure, we have assumed a reliability of .70, which under a normal distribution implies that virtually no true scores will appear to merit such remediation. With higher reliability, of course, more will be selected, and with perfect reliability, the true score will be identical with the observed score, and the same 2% will be selected in either case.

For selection purposes then, transformation to true score estimates does not appear to make any difference (though it might be a healthy practice in its effects on the attitudes of the personnel involved), when we are considering the case of a single test for a single program.

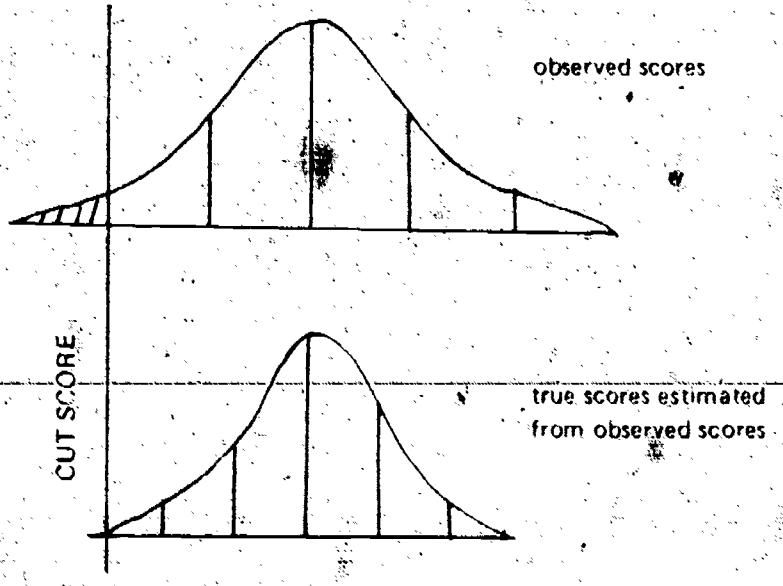
But one further case deserves notice: when we are limited (as implied by P.L. 94-142) to some quota for *all disabilities combined*. In such a case, where we are using cut-scores for inclusion, the nature of those included will be partly a function of the reliabilities of the selection instruments. For example: If in Figure 10(b), we had another test for another disability, and this second test had a reliability of .90, then we would include more students diagnosed by this second, more reliable test. In a sense, this would relegate to the *test*

224

FIGURE 10
CONTRAST OF TWO POLICIES
IN USE OF TRUE SCORE ESTIMATES



(a) A quota selection: the bottom 2%

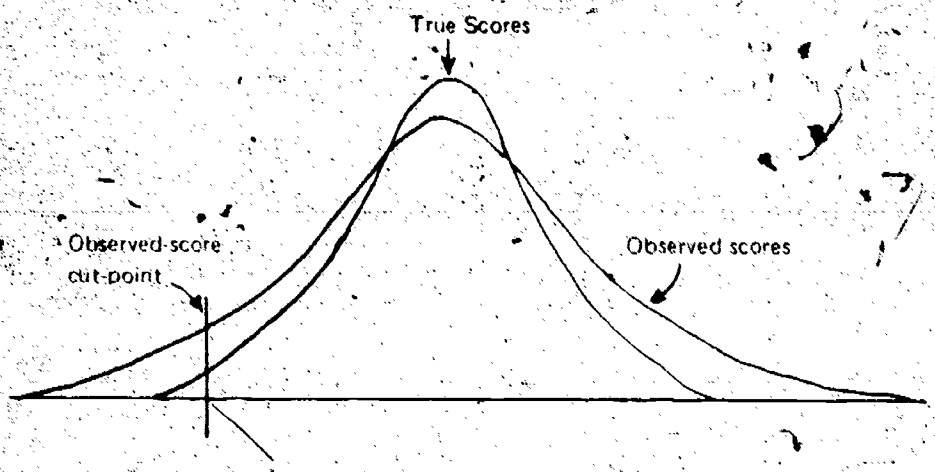


(b) A cut-point selection: those whose true scores are estimated below -2 standard deviations in observed units.

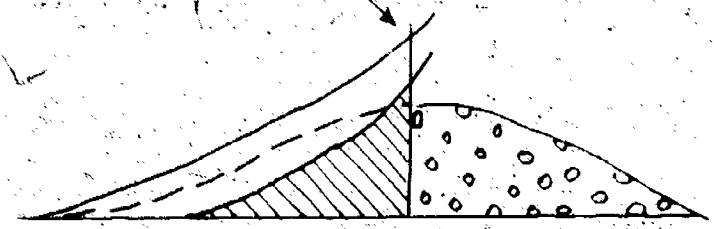
In case (a), exactly the same students are selected for the program. In case (b), almost no students are selected. There is an assumed reliability of .70.

225

FIGURE 11
TRUE HANDICAPPED AND
OBSERVED CUT-POINTS



(a) True-score distribution for an observed-score cut-point.



(b) Two types of error: Selecting "handicapped" who in true score are above cut-point; or failing to select handicapped who are in true score below the cut-point.

Depending on the reliability of the measure and the extremity of the cut-point, errors will be made of inclusion and exclusion.

characteristics some of the judgment more properly relegated to social and educational considerations. In this case, too, transforming to true score estimates creates as many problems as it seems to solve.

Reliability of Difference Scores

Originally, the guidelines for application of P.L. 94-142 had recommended a certain formula for calculation of a discrepancy score, in order to select those who were suitably below their "expected" level of performance in some important trait. These would be identified as fitting candidates for special remediation in those specific disabilities. Under much criticism from professionals, the formula was abandoned, and will not be reviewed here. But the problem remains with us: how to select among those otherwise "normal" youngsters the ones who genuinely are victims of such an LD. There are three sub-questions: How do we know the students are "normal"? How do we know they have a "specific disability"? And what is the meaning of any "discrepancy" between these two conditions? All are fraught with problems of unreliability.

As pointed out by others in this context (Ysseldyke, 1977; Salvia & Clark, 1972), the meaning of a difference between two scores is dependent on: the correlation between the two tests; the reliabilities of each; their standard deviations; and any differences between the original samples used for the norming. If we convert our school data of interest to z-scores, then we are still concerned with the three coefficients, and are anxious to know how much trust we may put in the difference,

$$D = Z_1 - Z_2$$

Suppose we ordain a policy that we are including only those who have an achievement 1.5σ below their achievement: Include student i if $D_i \geq 1.5$. What can we expect the distribution of D to be? In general, the variance of a difference is going to be

$$\sigma_d^2 = \sigma_1^2 + \sigma_2^2 - 2\,Cor(1, 2)$$

Where we have Z-scores, and $\sigma_1^2 = \sigma_2^2 = 1$, then this formula becomes

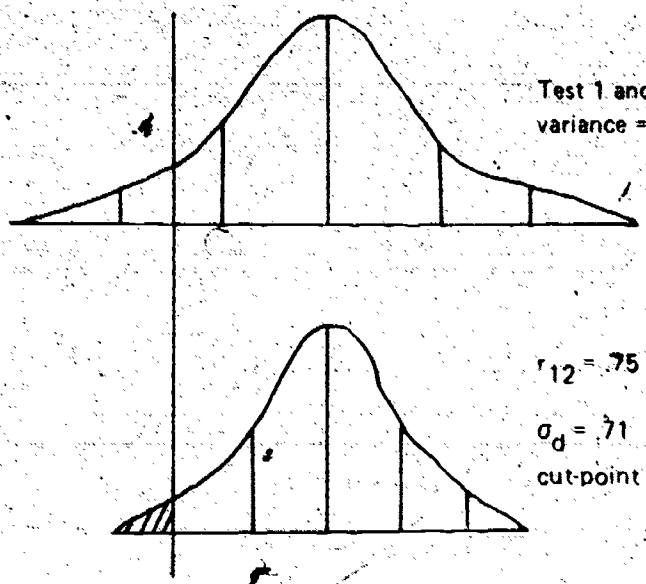
$$\sigma_d^2 = 2 - 2r_{12} = 2(1 - r_{12}).$$

If the correlation between achievement and potential is in the moderate range, say .50, then, $\sigma_d^2 = 1.00 = \sigma_d$. So, we would expect a cut-point of 1.5 to act as it would for a Z-score of a normal curve, and include .0668 of the population. On the other hand, if the correlation between achievement and potential is higher, say .75, then $\sigma_d^2 = .5$, and $\sigma_d = \sqrt{.5} = .71$. As shown in Figure 12(a), the result is that there is a cut-point for the observed difference score at $Z_{diff} = -2.11$, which makes a very conservative selection (under the normal curve) of .017 of the population. In general, then, the number selected for a difference cut-off will be a function of the correlation between the two measures. For comparison,

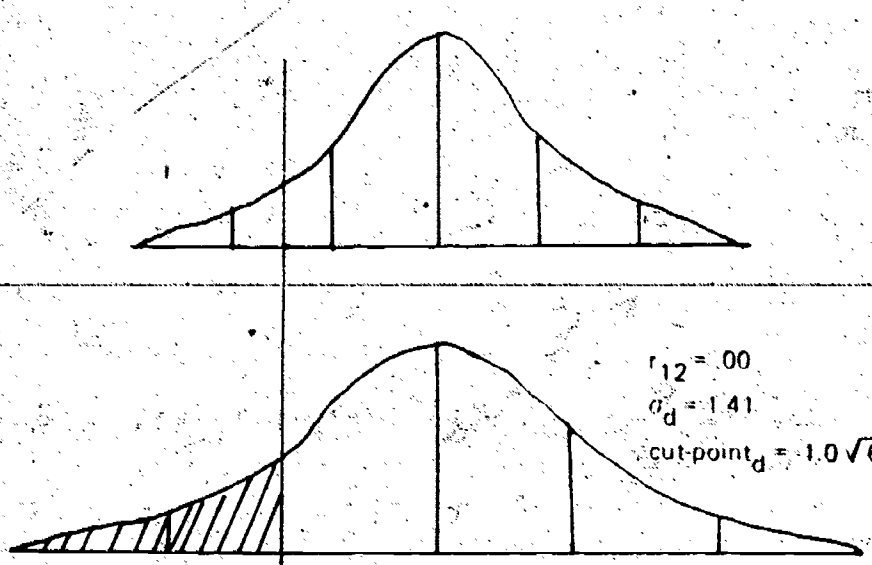
227

FIGURE 12

THE RELATION OF TEST INTERCORRELATION TO THE SELECTION PERCENTAGE FOR A DIFFERENCE SCORE



(a) When r = .75, just 1.7% of population is selected, when cut-point (in terms of test) = -1.5.



(b) When r = .00, then 14.5% of the population is selected, using the same cut-point as for (a).

consider the case of Figure 12(b), where we have a hypothetical correlation of zero between the measures. In this case, the cut-point of the difference score will be -1.06, for a selection of a substantial .145 of the population. (This is in contrast to the usual belief, and often comes as a shock to the person who is unfamiliar with such phenomena.) In contrast, the variance of difference scores, when r = 1.00, would be zero.

How may we best summarize the question of reliability of difference score? As noted by Ysseldyke,

$$r_{diff} = \frac{\frac{r_{11} + r_{22}}{2} - r_{12}}{1 - r_{12}}$$

The implication of this formula is clearly seen in a table from Stanley (1971, p.387), here reproduced for its meaning for selection.

TABLE 1
RELIABILITY OF A DIFFERENCE SCORE

Coefficient of Correlation Between the Two Tests	Mean Reliability Coefficient of the Two Tests $(r_{11} + r_{22}) / 2$					
	.50	.60	.70	.80	.90	.95
.95						.00
.90					.00	.50
.80				.00	.50	.75
.70			.00	.33	.67	.83
.60		.00	.25	.50	.75	.88
.50	.00	.20	.40	.60	.80	.90
.40	.17	.33	.50	.67	.83	.92
.30	.29	.43	.57	.71	.86	.93
.20	.38	.50	.62	.75	.88	.94
.10	.44	.56	.67	.78	.89	.94
.00	.50	.60	.70	.80	.90	.95

From Stanley (1971), p. 387.

229

This useful table reinforces many important ideas about this relationship. In the first place, the blank parts of the table occur where the correlation between the two tests (r_{12}) would be higher than the average reliability of the tests ($[r_{11} + r_{22}]/2$). Except for chance sampling variations, errors of estimate in the coefficients concerned, such an event cannot occur: therefore the blank. Next to the blank area is a diagonal of zeroes, for the case where the correlation between tests is just equal to the reliabilities of the tests. It must be intuitively apparent that, when this is the case, there is no *difference* between the two tests in what they are measuring; Test 1 is "the same as" Test 2. Thus there can be no real meaning, except error of measurement, in any *individual* difference between the score on one test vs. the score on the other: therefore the zeroes.

It is evident, then, that the reliability of the differences rapidly increases as the reliability of the two tests rises *vis-a-vis* the correlation between them. And the reliability of the difference would be perfect when the test reliabilities were themselves perfect, but the correlation between tests was zero.

Just where, in this table, is the typical situation where we would be selecting for learning disabilities, and controlling for potential? Note that any word we use for the control variable – "potential" or "expectancy" or predictor" or "aptitude" or "ability" – betrays the fact that there is normally a substantial correlation between that control variable (IQ or other measure) and the specific trait or skill for which we are selecting our LD children. Therefore, we would place our circumstances in a middling *row* of the Stanley table.

Now we consider the diagnostic LD tests, and examine a useful summary of their own reliabilities, seen in Table 2. In this Table, we observe that most of the frequently used tests have themselves reliabilities in the moderate range. If these specific tests are correlated with the ability measures, also in the moderate range, then we would expect, in the Stanley table, reliabilities for differences hovering in a very unfortunate area close to the diagonal of zeroes. The situation would be partly saved, depending on the reliability of the predictors – which, as in the case of IQ measures, can often be quite high. Yet we must face as well the problems created by part-whole relationships. To the extent that "Verbal Expression," for instance, will influence scores on a youngster's IQ test, we have a confounding of the two, since VE may be thought of as *part* of IQ, and the correlation between them will be raised, creating further difficulties in appraising the difference score. (And this is surely what happens with such a battery as the tabled *Illinois Test of Psycholinguistic Abilities.*) Beyond the purely psychometric problems, there are obviously philosophical questions in these part-whole relationships: when we "control for ability," do we *mean* the ability *without* the confounding specific disability? Serious discussions of this question are not easy to come by; and solutions are rarer still.

TABLE 2
TEST-RETEST RELIABILITIES
OF FREQUENTLY USED ABILITY MEASURES

Measure	Reported Test-Retest Reliability
Developmental Test of Visual Perception	.69
Eye Motor Coordination	.29 - .39
Figure Ground	.33 - .39
Form Constancy	.67 - .74
Position in Space	.35 - .70
Spatial Relations	.52 - .67
Bender Visual Motor Gestalt Test	.39 - .66
Chicago Test of Visual Discrimination	.35 - .68
Revised Visual Retention Test	.85
Memory for Designs Test (Graham Kendall)	.72 - .90
Primary Visual Motor Test	.82
Developmental Test of Visual Motor Integration	.80 - .87
Illinois Test of Psycholinguistic Abilities	.66 - .91
Auditory Receptions	.36 - .79
Visual Reception	.21 - .69
Auditory Association	.62 - .90
Visual Association	.32 - .75
Verbal Expression	.45 - .74
Manual Expression	.40 - .70
Grammatic Closure	.49 - .87
Visual Closure	.57 - .82
Auditory Sequential Memory	.61 - .89
Visual Sequential Memory	.12 - .71

From Ysseldyke and Salvia (1974).

231

Using True Scores with Test Differences

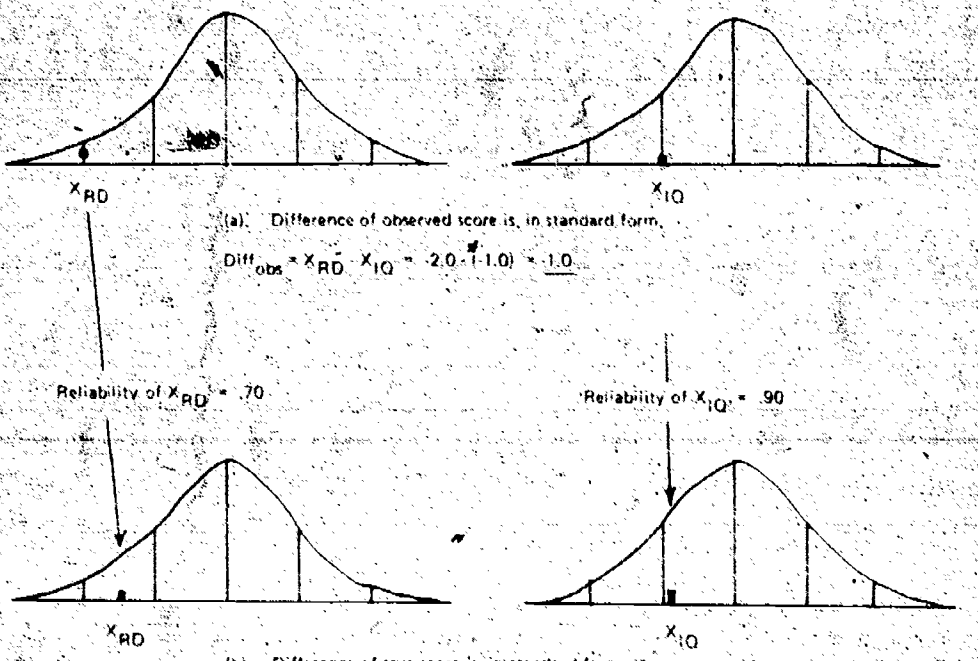
Earlier we examined the effects of using regressed true scores for specific tests of disability, and found the following: that since the transformation is linear, the rank order of selectees remains unvarying; that, therefore, when quotas are established (e.g., the bottom 2%) there is no difference between using true scores or observed scores; and that, when cut-points are established in terms of observed distributions, the selection is simply much more conservative for true scores (depending on the reliabilities of the tests used). In summary, the regressed true scores did not change the position of any subject in priority of selection, but with a cut-point policy, might greatly affect the *number* of selectees.

For difference scores, the recommendation of using true scores is also made. For example, Ysseldyke (1977) states: "I believe we should be computing reliabilities for differences between *estimated true scores* and using this information in making identification and placement decisions" [emphasis in the original]. His principal motivation, I believe, is to help the workers in the field gain a more careful and conservative attitude toward the differences reported, and this goal I strongly support. The only liability from this perspective of shaping attitudes, is that the term "true" itself might assume a completely spurious mantle of accuracy. The *true* score is, by definition, 100% accurate, but we never see it. The regressed, *estimated* true score, indeed, is closer to the probably true measure than is the observed. But, as we have seen, it is no more accurate in rank-ordering the pupils than that fallible observed score on which it is based. In fact, it rank-orders them the same way. In terms of professional use, then, it would be helpful if all the connotations of the word "true" were not carried along as excess baggage when we talked about these scores. Perhaps "regressed score" would be less misleading, but this would need some discussion, and possibly field tryout, to be sure.

However, it is a separate question what effects the estimated true scores would have on the tasks of making identification and placement decisions. For any difference score, $D_i = X_{i1} - Z_{i2}$, we have the same principles in operation as for single scores. In Figure 13, we see the two cases compared for difference scores. We presume for the Figure that we are comparing a Reading Score (X_{RD}) with an intelligence measure (X_{IQ}), with respective reliabilities of .70 and .90. We examine the case of a student whose Reading is -2.00 σ and whose IQ is -1.00 σ. In Case (a), then, the Difference Score is -1.00. In Case (b), we have regressed each score before computing the difference, with the result that the "true" Difference Score is now -.5 – just half what we found before. The effect of the true-score transformation here is to shrink the difference score (compared with that computed for the observed scores). However, if we give the higher reliability to the X_{RD}, and the lower to the X_{IQ}, we find that the "true" difference becomes slightly larger than the observed difference, -1.10. This happens because the reliabilitites serve as weights in the equation.

FIGURE 13

OBSERVED VS. TRUE DIFFERENCE SCORES



(a). Difference of observed score is, in standard form,

$$Diff_{obs} = X_{RD} - X_{IQ} = -2.0 - (-1.0) = -1.0$$

(b). Difference of true score is, in standard form,

$$Diff_{true} = X'_{RD} - X'_{IQ} = -1.4 - (-.9) = -.5$$

Thus we see that "deviation" scores, computed by such subtraction, are much more conservative in the number of subjects designated as discrepant, *if* some standard discrepancy is required for LD selection.

To see the effects of such reliabilities, I have generated the small Table 3, displaying the "true" differences for the above figure, under the assumption that the observed scores were the same, -2 and -1. Some of the tabled values for True Differences are a bit startling. Remembering that the observed reading score was a full standard deviation below the intelligence, what are we to make of a transformation which actually makes it appear *better* than the intelligence? Yet this is the case when we assume that $r_{11} = .25$, and $r_{22} = .75$. Granted that this is an extreme case, yet not an outlandish one, and this is for a substantial observed deficit in reading. Where the observed deficit is less, the reliabilities may be correspondingly closer to each other, and still produce such anomalies. Given the usual case where the ability scores will be more reliable than the specific scores, we can expect such apparent reversals of apparent "deficit" to occur again and again.

It is interesting to think through some of the algebra of the estimated difference score. Let us define

D = difference in observed Z-scores
D' = difference in estimated true Z-scores,
Z_1 = observed Z-score for a specific learning disability, and Z_1' is the corresponding true score,
Z_2 = observed Z-score for some control, ability measure, and Z_2' is the corresponding true score,
r_{11} = reliability of the LD measure, and
r_{22} = reliability of the control, ability measure.

Now we know that $Z_i' = r_{ii} Z_i$, and

$$D = Z_1 - Z_2$$

Thus $D' = Z_1' - Z_2' = r_{11} Z_1 - r_{22} Z_2$.

Let us assume, as in the typical case, that $Z_1 < Z_2 < 0$, and $r_{ii} > 0$. We can then work out the following facts about the difference score.

$D' = 0$ when $\frac{Z_1}{Z_2} = \frac{r_{22}}{r_{11}}$ (1)

$D' < 0$ when $\frac{Z_1}{Z_2} < \frac{r_{22}}{r_{11}}$ (2) and

$D' > 0$ when $\frac{Z_2}{Z_2} > \frac{r_{22}}{r_{11}}$ (3)

234

TABLE 3

DIFFERENCE IN ESTIMATED TRUE SCORES AS A FUNCTION OF BOTH RELIABILITIES

Reliability of Specific Score r_{11}	Reliability of Control Variable r_{22}			
	.25	.50	.75	1.00
.25	.25	.00	.25	.50
.50	-.75	.50	.25	.00
.75	-1.25	-1.00	-.75	-.50
1.00	-1.75	-1.50	-1.25	-1.00

Note: These values are all based on the situation of Figure 13, with a Specific Score of 2.00, and a Control Score of -1.00. All scores are in terms of the Z scores for the original measures.

These show that, when we compare the estimated true difference with zero, there is a regularity about the relationship between the two ratios, one for the reliabilities and the other for the ratios of the two tests.

Furthermore, we can compare the difference with the estimated true difference, and predict the relative sizes according to functions of these same ratios. We begin with the knowledge that

$$D - D' = (Z_1 - Z_2) - (r_{11}Z_1 - r_{22}Z_2),$$

and with the obvious statements of inequalities. From these we can derive, with the same assumptions as above, that:

$$D = D' \text{ when } \frac{Z_1}{Z_2} = \frac{1 - r_{22}}{1 - r_{11}} ; \quad (4)$$

$$D > D' \text{ when } \frac{Z_1}{Z_2} > \frac{1 - r_{22}}{1 - r_{11}} ; \quad (5) \text{ and}$$

$$D < D' \text{ when } \frac{Z_1}{Z_2} < \frac{1 - r_{22}}{1 - r_{11}} ; \quad (6)$$

And these equalities and inequalities are responsible for the behavior of these regressed scores.

These equalities and inequalities numbered (1) through (6) display the complexities we encounter when we transform scores to estimated true scores, and then attempt to understand better the differences between the potential and the achievement in the various deficiency areas. What keeps coming back to us is that regressing the true score, while it gives us a value closer to the *long-run* observed for a student, will not help us rank-order the students, in order of need or priority, any better than the observed score, exactly because it is equally filled with error. And as suggested, the label "true" may serve in the local agency further to inspire a misplaced confidence in the resulting numbers.

Regarding the use of the *reliabilities* of the "true differences," these would appear to be the same coefficients as those for the observed differences, since the same uncertainties enter in each case. However, the standard error of measurement of the difference (SEM_{diff}) would apparently be different, since the standard deviations of the regressed scores would clearly be different from those for the observed. Indeed, without working it out formally here, it seems that the (SEM_{diff}), if regularly reported and transformed into probability distributions about the estimated value, would play the same role as working with the true scores, so far as we are concerned with the proper display of uncertainty of classification. That is, the local agency could work with such classifications as the following: "Johnny's Reading seems to be one standard deviation below his General Ability. However, because of the reliabilities of the

236

tests concerned, there is only a 20% chance that this is really the case, and there is a 40% chance that his Reading is as high as his General Ability."

Such formulations are obviously not easy to work out at the local level, by any manual process. It is therefore suggested that computer programs be developed nationally to perform just this sort of printout, and that these programs and packages become easily available to the systems' data-processing specialists at the state and local levels. The programs would be easy to develop, whether under government or private auspices, and could be made quite general so that the characteristics of the component tests could be entered easily as input; or they could be made quite specific to the particular test-comparisons desired.

All of these complexities add more doubt about the roles played by many of the professionals serving on the CETs at the local level. How many will be able to recognize these problems of classification, and take proper account of them? How many professionals (at *any* level) could work out such individual probabilities at their desks, even if they had time to do so? Yet such complexities, however technical, are important in such cases, and not "soft" - that is, subject to overriding reinterpretation of a subjective sort. The reservations about "clinical" vs. "statistical" judgment are particularly relevant when the quantitative realities are as complex as these.

Increasing the Reliabilities

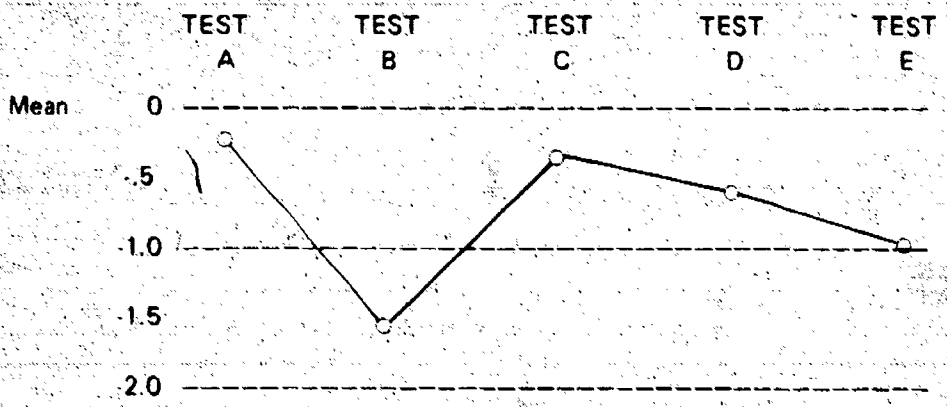
Recognizing the limitations of current strategies is one thing; but improving the reliabilities of the component tests is something else, and certain to be strongly desired. To the extent we can improve such reliabilities, then the true-vs.-observed-score issue becomes of less importance. Some possible methods of doing so are the following:

1) Do much more *testing*. Two tests will combined have considerably more reliability (and probably validity) than one alone. To the extent these may be done with less expensive testing (group testing or self-administered testing), or may be respectably performed by less-trained professionals, the cost may be reduced. But in any case, the uncertainties of classification, and the likelihood of wasting large amounts of money through such uncertainties, argue for much more of the total investment to be spent in proper selection, diagnosis, and prescription.

2) Use the most reliable *combinations* of data. What do we mean, in general, by "potential"? If we are talking about intelligence, then the most general, and really fundamental, definition of it is in terms of the pattern of positive intercorrelations found in a matrix of mental tests. This is Spearman's *g*, the first principal component of a battery of mental tests. The development of "IQ" tests

is largely a search for those items which correlate most generally with this *g* factor, which load most heavily on it in a first factor matrix, before rotation. Let us put this in visual terms, as in Figure 14.

FIGURE 14
IS THE PUPIL DISCREPANT IN TEST B?



Since the reliability of a battery is greater than that of a single test, the general factor of the other tests may be used to increase the sensitivity of the comparison with Test B.

Here we see that the profile of a student, from a battery of mental tests, may have some variation. In general, however, *g* will imply some similarity of test range. If our concern is with whether *B* is sufficiently lower to merit special treatment, we may increase the reliability of the comparison by using the battery as a whole. This may be done either through the loadings on the known general factor, or perhaps more conveniently through the "total scores" furnished by some test publishers. We can observe, in the Stanley table, that increasing the reliability of one test substantially increases the reliability of the difference score.

3) Increasing the *relevant* items. A decision system requires additional attention to the zone of the greatest uncertainty. Millman (1974) suggests constructing an uncertainty band around the passing standard (for any selection). Those with scores in the band would be given more items so that a more precise estimation of true scores may be made. As we have seen, such "uncertainty bands" might be quite wide. Uncertainty bands are also accepted by Swaminathan, Hambleton, and Algina (1975).

Agreement of Judges

In order to understand the diagnosis of Learning Disability, let us make some simplifying assumptions: 1) There is a known correlation betweenjudgments of LD for the population at large, which we shall call r_{AB}. 2) For our purposes, the distributions of diagnostic scores will be approximately normal. 3) For these same purposes, the regression of B on A (or A on B) is homoskedastic. These are commonplace assumptions in use of regression.

Under these, the concordance of judgment is much like Figure 15. Part (a) shows that portion of the normal curve chosen by A under the limitation of choosing just 2%. Figure 15(b) shows the new distribution of A's selections, as distributed by B, when A and B are correlated .60 with each other. Here we observe that *most* of A's choices are not selected as LD by B; in fact, only around one quarter of A's choices were so selected. (Chance selection, with no correlation between A and B, would result in one fiftieth common selection, since when A and B are independent, $P(B/A) = P(B)$.)

Now consider the case where $r_{AB} = .80$ – a very rare agreement among judges where there is a large subjective element. Even so, most of A's choices are unchosen by B, as shown in Figure 15(c). Even with such a high concordance, then, a school system would experience chaos in administering an appeals program, with the first diagnosis typically overturned by the first appeal.

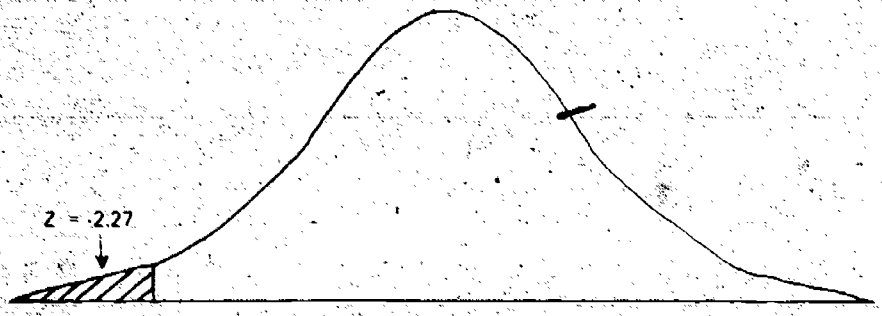
One other perspective on this question is exhibited in Figure 16. Given that Judge A made the first selection, and the correlation between A and B is .60, the Figure shows the small, cross-hatched portion to illustrate the concordance of selection into an extreme program, with 2% admitted.

In sum, the system of appeals appears to be in a dilemma: either the appeal level will be influenced by the original judgment (in which case it is not an independent appraisal), or it will not be so influenced (in which case it will quite possibly not decide the same way). If the appeal is from a *selection* for LD treatment, then the appeal level will typically overturn the original placement. If the appeal is against an *exclusion* from the LD program, then the appeal level will typically support the original decision. And the difference between these two cases is, as we have seen, an artifact of the unreliability of judgment.

Further Questions about Appeal

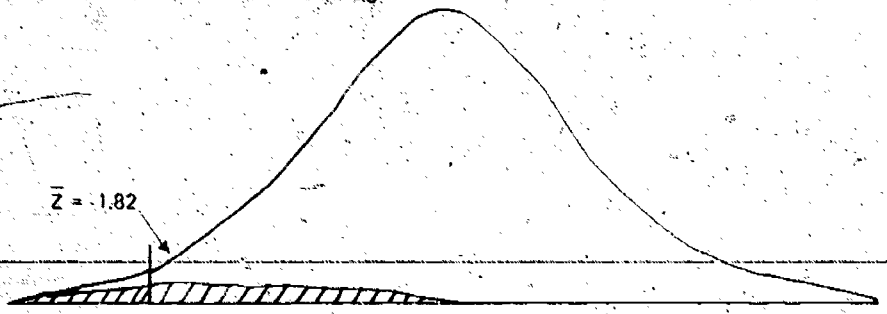
There is much that is uncertain about this second judgment. In general, it is well established that clinical judgment is less certain than more mechanical judgments from objective measures (Hills, 1971). One question, then, is whether the appeal judge will be aware of the former judgment made. If so, then any bias in Judge

FIGURE 15
THE SELECTION OF EXTREME CASES
BY TWO INDEPENDENT JUDGES



(a) Shaded portion is LDs chosen by Judge A as "bottom 2%."

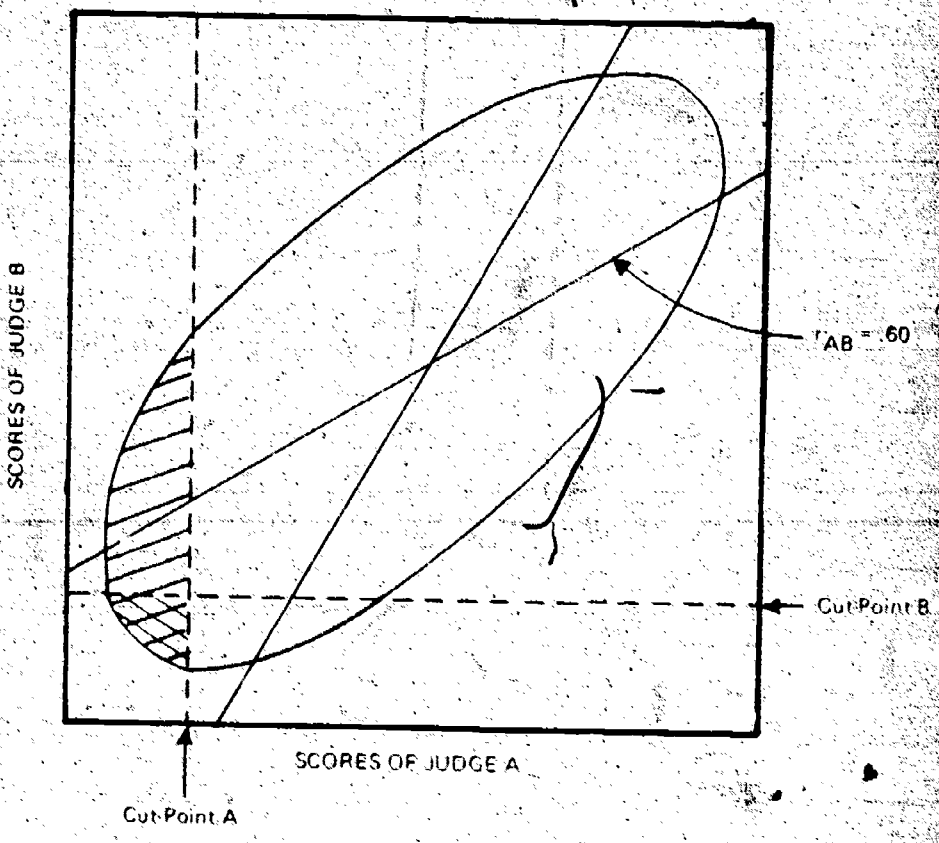
(b) Shaded portion is distribution of A's choices as scored by Judge B, given a correlation $r_{AB} = .60$.



(c) Shaded portion is distribution of A's choices as scored by Judge B, given a correlation $r_{AB} = .80$.

In (b), only one-quarter of A's choices are chosen by B. In (c), only two fifths of A's choices are seconded by B. (See Note 1 at end of chapter.)

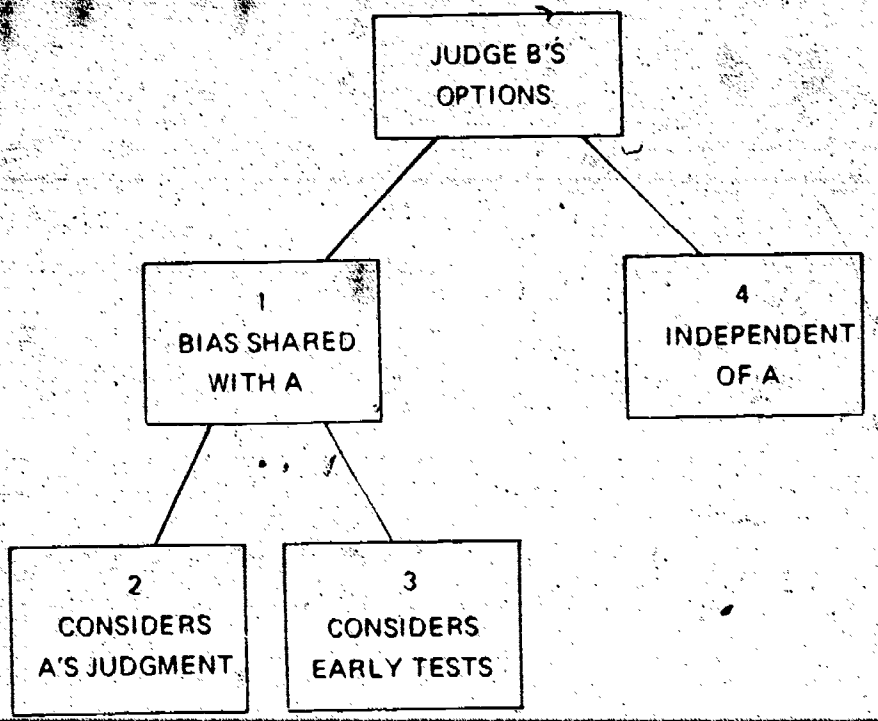
FIGURE 16



The concordance of two independent judges, with r_{AB} = .60, in selecting an extreme group. The probability of A/B is about .25. (See note 2 at end of chapter.)

A's decision will have a possibility of contaminating Judge B's decision. If not, then we have the question of whether Judge B will look at the same objective *test* as Judge A (or whether such tests will be repeated independently). If Judge B *does* consider the same tests, then his judgment, once again, will share any bias due to the first test. If not, if Judge B somehow starts fresh, then we have a fairly well-understood situation, but not an encouraging one. These possibilities are seen in Figure 17, where a tree is drawn of some of Judge B's possible actions.

FIGURE 17
THE DILEMA OF A SECOND JUDGEMENT



Either (1) Judge B shares bias of Judge A through either (2) considering A's judgment, or (3) considering the same tests and other materials used by A. *Or* (4) Judge B acts independently of A (without knowledge of A). In cases 2 and 3, there is indeterminate bias. In case 4, concordance of assignment is highly unlikely, as explained in text.

If Judge B has information about Judge A's actions, then the bias (if any) in Judge A's decision will contaminate Judge B to an unknown degree. If Judge B operates "blind," without knowing Judge A's prior decision, then we may consider Judge B "independent," and any common bias would be the result of common biasing principles, rather than of contamination of B by A. If Judge B is indeed independent of A, then what may we expect in terms of agreement?

The Generality of this Difficulty

Note that, so far as Figure 17 is concerned, one "judge" is like another, whether the judges be individual human decision-makers, or committees charged with the decision, or tests administered for the same decision-making purpose. The operating parameter is the size of the correlation between decision-makers. Since more objective measures generally correlate more highly with each other, we may ask about the built-in difficulty where such agreement is like a good reliability, about .90. In such a case, we reach at last a situation where most of A's choices are also chosen by B — but just barely, since around 46% of A's choices would be unchosen by B.

If we assume that there are, indeed, many more than 2% who "need" special attention, then this difficulty of choosing extreme cases reliably is not too severe: If *anyone* in the bottom two standard deviations (16%) needs help, then no particular damage is done by not selecting exactly the bottom 2%. For the purposes of selection into the program, then, we may consider the error not too malign.

The great disturbance, as we have seen, comes with the expectation that *appeal* will provide a system for redressing wrong. It will do no such thing, for it is caught in the dilemma: (1) of being guided by the previous fallible decision, in which case it will share the same biases and errors of judgment, or (2) of being independent, in which case it will mean a nearly systematic overturning of the earlier judgment — placing an enormous strain on the system. In this dilemma, there is no such thing as a "compromise": If Judge B gives *any* attention to the earlier judgment, he/she/they/it will be contaminated to an undetermined degree.

This chapter has concentrated on questions of reliability: in tests, in difference scores, in agreement of judges, in the use of appeal procedures. We have noted some recommendations about improving reliabilities, but the overall recommendations for the local and state systems are saved until the summary of Chapter IV.

Notes on Figures 15 and 16.

Note 1

For Figure 15 (b, c), the overlap was estimated as follows: The mean of the bottom 2% of the normal curve has a Z-score of about -2.27. The mean of the predicted curve (for Judge B) will fall at about $r_{AB} \times 2.27$. Thus for Figure 15(b), the mean of predictions is taken to be $2.27 \times .6 = -1.36$. The error in estimation for a single predicted point is $\sqrt{1 - r^2}$, which for $r = .60$ yields an error deviation of .80. For a range of such predictions, the error would be slightly larger, making a slightly more favorable outcome.

Note 2

The degree of contamination of Judge B by the materials of Judge A or by Judge A's own judgment, may be researched, as follows: Investigate the three situations as shown in Figure 16. For each of the cases, collect information about the concordance of decisions of Judge A and Judge B. Case 3 gives us the independent measure of agreement between judges, r_3. The observed correlation for Case 1, where Judge B is aware of Judge A's decision, will be r_1. The influence of Judge A's prior decision, then, can be estimated by $r_1 - r_3$. If we call such influence b_A, and assume that this is independent of r_3, then, $b_A = \sqrt{r_1^2 - r_3^3}$ after the correlations are transformed to their z-score equivalents.

CHAPTER III:
SPECIAL PROBLEMS OF FAIRNESS FOR HANDICAPPED

Are Treatments for Handicapped Effective?

As we have noted, "fairness" of testing depends on what use is made of the test results; and the fairness of the use depends on such factors as a student's probability of showing benefit from the prescribed treatment. To receive a given score, a given classification, or even a given treatment, is not in itself right or wrong, fair or unfair: but to receive (within the resources available) treatments which will alleviate the handicap – this may be considered a criterion of fairness to the individual.

Yet often, to raise the question of effectiveness of treatment seems to put a critic in the role of the skeleton at the feast. P.L. 94-142 was obviously designed under the impression that we had effective diagnostic and prescriptive competencies, and treatments of proved efficacy. Ysseldyke (1977) forthrightly states there is little "empirical evidence to support the contention that specific interventions or treatments lead to desirable academic outcomes," and again raises the issue of "the extent to which we can continue to assign students to instructional interventions with little if any empirical support for the efficacy of those interventions" (see also Ysseldyke & Salvia, 1974).

Clearly, we need far more evaluations of programs for the handicapped; and these must somehow escape the problems of the past. Above all, evaluations should be conducted by competent, skeptical, and critical outsiders – not those who have much to benefit by the finding of successful results. The past twenty years have demonstrated how often we may be confused by results of ad hoc programs, without standardized evaluation techniques or external controls (Page, 1972a). The greatest insights seem to come when external agencies survey material competently across substantial data sets (Colemen et al, 1966; Page, 1972d).

Yet even large and capable researches will often fail when working with a very poorly understood response surface. Ysseldyke (1977) has correctly identified two themes in special education: one concerned with causes of the disabilities; the other arguing that causes are quite beside the point, provided the "specific disabilities" can be corrected. One of these is more akin to cognitive psychology, the other to behavior mod. One is more akin to traditional science, the other more to practical engineering. One can surely argue both sides.

Yet in the long run, however much we need specific and immediate remedies, we must keep pushing for better understanding of the causes. A key question about

245

handicap, for example, is the degree to which it is genetic in origin. And we will pay some brief attention to this question.

Heritability of Specific Abilities

Heritability, to h^2, is defined as the proportion of the variance of a trait which is attributable genetic variation. Quantitative geneticists (Falconer, 1960; Mather & Jinks, 1971) and behavioral geneticists (McClearn & DeFries, 1973) are frequently concerned with separating different kinds of genetic variance. But most educators and psychologists are more concerned with *broad* heritability, the total degree to which heredity accounts for measured behavior – exactly because, professionally speaking, we are not concerned with heredity at all. That is, if the total variance of a trait consists of

$$\text{variance} = h^2 + e^2 = 1$$

then as educators we are concerned with $1 - h^2$, that variance remaining within the control of the environment. The implications of h^2, then, are interesting regarding what we may hope to change. (Cf. Bereiter, 1970.)

In general, the evidence of high, broad heritability of mental ability is now so consensual that it finds its way into standard elementary textbooks (e.g., Hilgard, Atkinson, & Atkinson, 1975) without debate. What is more interesting and much less understood is the heritability of profile differences, of specific abilities. Let us consider the list of certain prescribed Learning Disabilities under P.L. 94-142: oral expression, listening comprehension, written expression, basic reading skill, reading comprehension, mathematics calculation, mathematics reasoning. When we find marked profile differences in such skills, to what should we attribute them? To genetic differences in the traits? To past environmental histories? To errors of measurement? As we shall see, the concern about errors of measurement is a real one. But once reliable profiles are established, what may one say about the remaining differences between the true scores?

If h^2 were 100% of the variance in true scores, then we would have little hope of altering the trait. (That is, by environmental differences such as found in our culture. In theory, if we had some effective treatment available which was *hardly ever found in our culture*, we might hope to improve the trait, regardless of the *present* high heritability.) On the other hand, if h^2 were very low for the true scores, then we could hope for various "equalization" programs having a large effect on the trait, so long as we could identify the environmental variables which were responsible for the trait variance.

So far as I know, not much attention has been given to the specific Learning Disabilities, regarding their heritability. Some related information, however, is available to us from other work. The usual procedure in large testing programs is

to identify pairs of twins in the population, classify them according to zygosity, then calculate h^2 according to certain assumptions (which there is no space for here). One large data set was generated by the National Merit Scholarship Qualifying Tests of the early 1960's, and interestingly analyzed by Loehlin and Nichols (1976). From the intraclass correlations they present, we can estimate h^2 for a number of traits obviously related to the Learning Disabilities of the Law. We show such estimates of h^2, and the accompanying reliabilities of the traits in Table 4.

TABLE 4
HERITABILITIES AND RELIABILITIES OF FIVE ACHIEVEMENT MEASURES FROM THE NATIONAL MERIT TWIN SAMPLE

(N of pairs = 850)

Measure	h^2	r_{xx}
English Usage	.38	.91
Mathematics	.49	.88
Social Studies	.43	.84
Natural Science	.35	.83
Vocabulary	.48	.96

The tabled values of h^2 have not been corrected for the unreliability of the scores, an alteration that would typically increase the recorded heritability. But these are only indications from a very recent study. As shown elsewhere, the heritabilities of scholastic achievement have a fair range of moderate coefficients in the literature (e.g., Jensen, 1973, Ch. 4). In Australia, investigators found heritabilities for a long list of achievement tests for 15-year-olds, and they were as high as those for IQ, largely in the .70's and .80's (Martin, 1975, p. 225). Still, one could argue that the heritability of achievements is determined by the loading on a heritable *g* factor; that the profile *differences*, in other words, were the result of environment.

To test this heritability of specific traits, Martin (1975) looked for telltale interactions and covariances for different school achievements. With small sample sizes, most tests were not conclusive. Still, he found what appear to be different trait loadings on different gene loci for comparisons of English and Mathematics, and of IQ and height ($p < .01$). This study only suggests techniques for analysis of such questions, and points to a fair probability of finding specific

heritabilities for the traits we speak of as Learning Disabilities. Beyond any such demonstration of existence, however, is the more important question of the amounts of specific loadings on genetic or selected environmental factors. For trait j, in other words, just how much of the variance is environmental variance *specific* to trait j?

It is urged, therefore, that evaluation be promoted in three ways: specific local analyses of results, large-scale summative evaluations of different types of programs and their apparent effectiveness; and more basic researches of the sources of the traits, through techniques of causal path analysis, dealing with both environmental and genetic influences. Apart from the h^2 estimates, it is also desirable to study siblings for environmental influences between and within families.

Such more basic research will not immediately affect local programs; but at a national level, as evidence is gathered in, we should be able to make better predictions about the effectiveness of interventions in various LD areas. This would provide us with the "production functions" needed to use in more sophisticated decision-making formulations.

Cultural Bias and Assessment

It is sometimes reasoned that learning disabilities should be judged differently for different ethnic groups. That is to say, mental retardation would be indicated as (for example) two standard deviations below the group of the subject child. And LD conditions would be diagnosed less often, for members of groups which were, on the average, lower in the skill concerned. A clear statement of this position is from Mercer (1972):

> Finally, in a pluralistic assessment, the meaning of a particular test score or adaptive behavior score should be interpreted not only within the framework of the standardized norms based on a sample of Anglo children [by this is meant the white, English-speaking majority] but should also be evaluated in relation to the sociocultural group to which the child belongs. . . . His position on the norms for his own sociocultural group indicates his probable potential for learning (p. 445)

In this part, let us consider briefly some arguments about this matter, in the case of native-born, English-speaking children of whatever ethnic or national background. As commonly interpreted, the above recommendation is believed applicable to Black-White differences, but is seldom proposed for any other English-speaking groups, so it is in this context it will be considered.

There is little doubt that the Black-White difference is a cause of real concern. Perhaps the most comprehensive investigation of IQ differences (Shuey, 1966) summarized the race and class interaction as seen in Table 5. The striking thing

TABLE 5

AVERAGE IQ SCORES BY RACE AND BY SOCIAL CLASS

Race	Social class: Upper	Lower	Difference
White	111.88	94.22	17.66
Negro	91.63	82.04	9.59
Difference	20.25	12.18	

Data from Shuey (1966), p. 520.

here is not the class differences, or the race differences, but rather the large race differences within the social class, and especially the fact that, across the hundreds of studies summarized in the table, the lower-class White group exceeded the higher-class Negro group. This is not an unusual finding, though it remains a finding that is ignored in much planning. It was also found, in quite different samples, by Coleman *et al.* (1966), Wilson (1967), Scarr-Salapatek (1971), and Page and Grandon (1977). In this last, the large samples from the National Logitudinal Study were divided into three social classes: the top 25%, middle 50%, and bottom 25%, on the basis of five variables of income, parental occupation, and parental education. The same cut-points were used for SES for all students. Here, as in the other studies, the top-SES mean for the Negros fell just below the bottom-SES mean for the Whites. (This finding was a by-product of the search for explanations of family-size effects.)

This finding of persisting Black-White differences despite controlling for social class is of course the center of the "nature-nurture debate," as widely waged in the social sciences and education. It is clear, from such means as those tabled, that no regression will eliminate the race effect, unless some variable is used which is in fact a functional equivalent for race. Such a variable, really serving as race identifier by another name, is residence in a community that is black, or school attendance in a predominantly black school. When this is partialled out of a regression of intelligence on race, it essentially removes the variance attributable to race, and consequently the means may be considered "equal." But otherwise, it seems impossible to explain the race difference using the best measures we otherwise employ to study cultural advantage. The current debates center around explanations which are either completely environmental, or which provide for some genetic influences in the race difference.

This is no place to review any of these arguments, which have absorbed already many books (Cancro, 1971; Jensen, 1973; Lodehlin, Lindzey, & Spuhler, 1975; Block & Dworkin, 1976; Hebert, 1977). The bottom line, however, must continue to loom large in discussions of correct procedures for "classification" of MRs or LDs, or for that matter of gifted or talented (Page, 1976a.). Important for some writers seems to be the "labeling" itself: that is, it is not what is done with the child which is believed damaging so much as what the child is *called* in order to diagnose and prescribe. There is no doubt that, if a color-blind MR or LD classification is used, there will continue to be vastly disproportionate numbers of Black youngsters included; and it should be noted, by examining the table above, that a substantial number of these included Blacks will be from families which are *middle- or upper-class*, a situation which undoubtedly creates considerable and unusual strain in relations with the aspiring, well-educated, and affluent Black parents.

Therefore, we should ask what conclusions are currently justified by the available evidence and logic. On the question of geneticity, one of the most recent, thorough, and objective analyses is forced to a still-uncertain "outcome consistant with a genetic interpretation but not necessarily excluding a subcultural one" (Loehlin, Lindzey, & Spuhler, 1975, p. 238). These authors, like Nichols (1978), outline a number of promising avenues of research into the question for the near future, including the eugenics-dysgenics trends, which for minorities they term as "rather more urgent" (p. 256) than for the society as a whole. Scientifically, then, if we review carefully the status of the question among most scientists who have studied the issues, we must conclude that there is a clear consensus that genes are important in intelligence, but not a clear consensus about the possible genetic contributions to the widely observed race differences in intelligence and school achievement.

If the jury is still out scientifically, it is also still out legally. The term "reverse discrimination" has possibly overtaken the term "affirmative action" in the press, in stories centering around a number of court tests of racially selective programs (e.g., *Bakke v. the University of California*). And the question of whether tests are fair to Blacks is the subject of other court trials (e.g., *Larry P. v. Wilson Riles*).

It is necessary to stress this very incomplete situation, both scientifically and legally, when we consider the possible use of tests for the handicapped. If we set aside the nature-nurture controversy, and study only the relationships outside the genetic questions, then the consensus of scientists really becomes quite decisive: there is not real "bias" in current tests, if we take as evidence the vast body of accumulated data about the relation of tests to other tests, and to other measures in the educational and vocational world. Investigators quite outside the genetic debate have concluded that the races share essentially the same regression line (for the prediction of other performance through testing) for

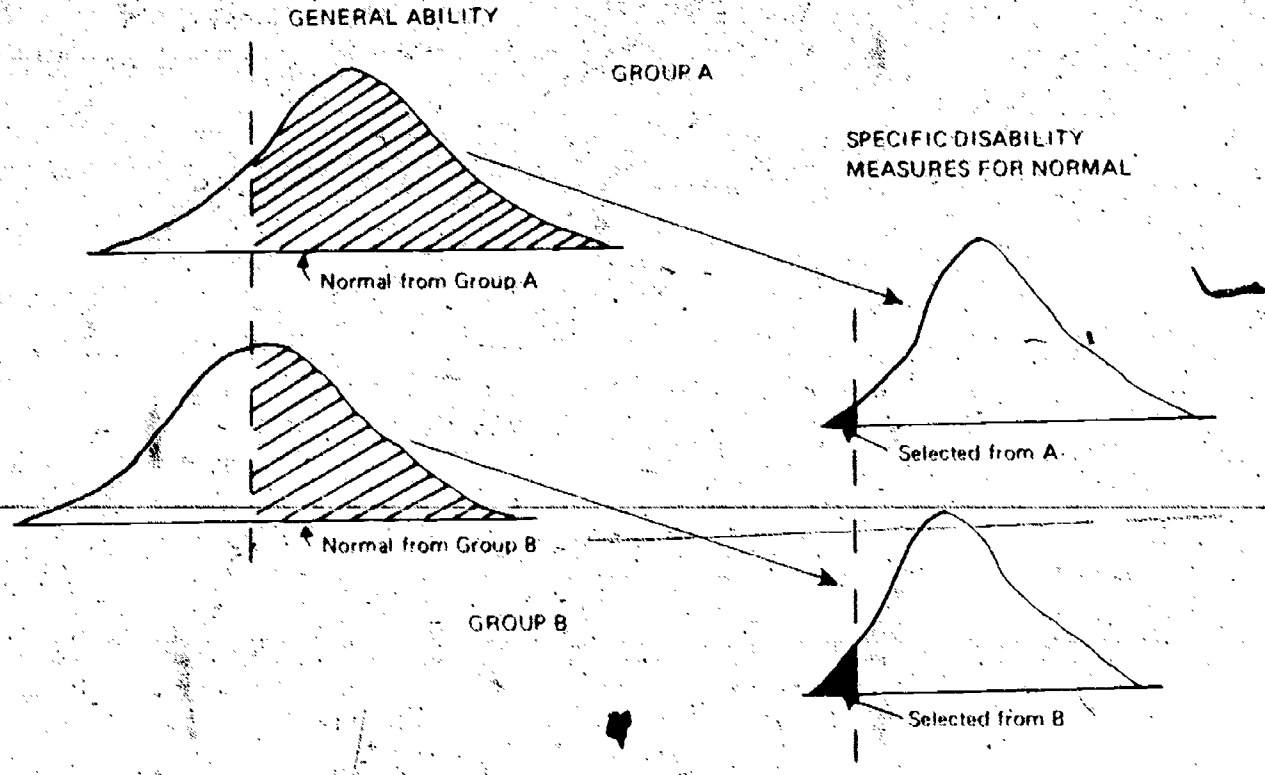
most of the important uses (e.g., Cleary, Humphreys, Kendrick, & Wesman, 1975; Stanley, 1971a). It is still possible, of course, to claim that these criterion measures (success in college, vocations, etc.) are themselves biased. But it is apparently impossible to support such a claim, apart from appealing to the observed differences themselves, which would beg the question at issue and lead us into a circle.

Psychometrically, then, I believe we must conclude that, for English-speaking minorities, whether Black, Oriental, Jewish, Eastern European, or other, the most useful policy, within what we know of testing and what testing predicts, and what programs and procedures are available and useful for remediation — for the current situation as a whole — we should probably follow a purely individual policy of testing for handicap. This implies that ethnic considerations (apart from foreign-language speakers) should probably not play a role. Note that this judgment *is* a psychometric one, and is not intended to prejudge the court tests, nor to settle political issues raised by groups seeking greater justice or power.

What would be the practical effect of such a color-blind policy? It would mean that the classification of children as the victims of "sociocultural" injustice would become less frequent. It would mean that the proportion of Blacks classified as MRs (or some other more palatable term describing the same condition) would continue to be large, and in fact perhaps grow. But the effect on the composition of LD programs would depend on some unresolved questions: The Federal guidelines define LD only as a "severe discrepancy" between ability and achievement, and do not further define "severe discrepancy." Therefore, the LEA must operationally define LD for itself. If LD is defined in terms of absolute handicap relative to the population, then the proportion of Blacks would be large. If LD is defined in terms of deficit *within an individual profile* (i.e., achievement after controlling for "potential"), then the proportion of LDs would probably be not much different whatever the ethnic group considered. (This is another way of saying that the variability within profiles probably does not greatly differ from group to group.) The most probable resolution of the guidelines will probably be a combination of these two considerations, a compromise between these extremes, selecting those quite low in LD category, but within a "normal" range in general ability. To see how such a system might select LD youngsters, we include Figure 18.

In Figure 18, we construct two populations *A* and *B*, to be loosely analogous to the White and Black populations. (The letters are used to emphasize that the figures do not represent data, but theoretical reasoning.) Group *A* has one σ higher "general ability" than Group *B*, as measured either by IQ tests or by some *g* from battery profiles. According to the assumptions here, only those members of each group in the "normal" range are eligible for consideration for LD selection. Since the "normal" for Group A has a higher average than for Group B, we can expect some of this difference to be reflected on a specific measure of

FIGURE 18
POSSIBLE GROUP DIFFERENCES IN SELECTING LDS



If only those with "normal" general ability are considered for LD classification, the group with lower ability will still yield a larger proportion of LD classification.

some learning (dis)ability, which will be correlated with the general ability. Thus the normal *A* group, when tested for LD, forms a distribution indicated by the arrow from *A*: a group with a somewhat smaller σ and a slight skewness toward the high side (i.e., a blunting at the low end). Group *B* normals will project a similar distribution for the LD measure, except that their LD distribution will be slightly narrower still, and somewhat lower than that of Group *A*. The difference will no longer be one σ, since there will be "controlling" for general ability, but it will still exist. The net result is that there will continue to be an over-representation of *B* students in the LD category.

And although these drawings are not at all to scale, the inference is that, in the most probable resolution of the classification problem, *if color itself is not used as an exception*, there will continue to be a disproportionate number of Blacks assigned to programs in LD. Of course, if color *is* used as a criterion, then these psychometric comments would not apply. Then the judgment could be manipulated so that, for instance, Blacks would be only proportionately represented in any LD category, or in the handicapped program overall. But it should be recognized, before implementing such a color-criterion, that the MR and LD classifications will not mean the same thing in child behavior or in child learning for the two races. Still more evident will be the remaining differences in the "normal" group, which will include Blacks and Whites of quite different psychometric characteristics. Since Blacks and Whites share pretty much the same regression line, we could expect that the bottom achievers among the "normals" would be disproportionately Black, despite their having ability scores altered upward out of racial considerations.

Concerning such different ethnic norms, the most important question may be one we have not asked, and one commonly side-stepped in discussion of the Law: Do we really have skills and programs of proved effectiveness for the various MR and LD classifications, once the appropriate pupils are identified? If not, then such psychometric niceties as we have been describing become less important. If so, then it would appear to be a disservice to the Black children to deny them the benefits of such programs appropriate to their correct (psychometric) classification.

The Native Language Requirement and Assessment

The law calls for tests being "nondiscriminatory" to different language groups, and mandates that a test must be given in a child's "natural language." Though written in general terms, this phrase of the Law has a widely acknowledge target group: the "Hispanics", that group made up of "Latins," Puerto Ricans, Mexican Americans, and other Spanish-speaking young people, mostly from Central and South America, many of them belonging to migrant families. For

this is a group showing marked deficits both in school achievement, and in the various aptitude and ability tests intended to provide guidance for youngsters and evaluation of their development.

American immigrants have, of course, been of foreign-language background more often than not. For such groups, the school systems of the U.S. have typically been a shaping force for the second generation, insuring that the common language, English, becomes a workable tool for the immigrant's offspring, and otherwise pressing the newcomer toward an adaptive acculturation to the majority society. For a variety of recent sociological and political reasons, there has been an overriding trend to treat this one cluster of immigrant peoples, the Hispanics, as a special case; and recently many laws have been passed, and practices adopted, encouraging the continuation of Spanish language through the schools and lives of continuing generations. Whether this is helpful either to the Hispanics themselves or to the background culture, has in fact not been resolved, and even investigation of it seems remarkably scanty. But insuring that there be no language discrimination against the Hispanics is a governmental action that may be justified quite apart from the validity of the political movements from which it derives its impetus.

Consideration of discrimination and language, then, is most centrally related to the problems of Spanish-surnamed children. Much of the concern, speaking professionally, has probably stemmed from the writings of Jane Mercer (1972, 1977), and it seems sound to consider her principal arguments regarding the Hispanics in the districts researched in California.

1) There is the repeated evidence that the observed scores of the Chicanos are well below those of the majority children.

2) As a result, many more Chicanos are termed "mentally retarded" and are placed in special classes, than is true of the majority children.

3) Mercer believes there is an abuse of the Chicano children in "labeling" them as "MR" on the basis of English-language tests.

4) Mercer tested the hypothesis that the Chicanos might genetically have more MRs, by partialling out variables related to "sociocultural factors" on which there was a large difference between Chicanos and majority children. When all such factors (including Spanish-speaking parents) were partialled out, there was no important residual difference, between the two groups of children, in measured ability.

5) Mercer has called for "pluralistic assessment" of such children, providing for norms within the minority group itself, as a way of avoiding the mislabeling and the consequent assignment to inappropriate educational programs.

To consider this problem psychometrically, we shall consider the plight of the Hispanics in several regards: their patterns of mental test scores; the use of tests for diagnosis and assignment; the diagnoses as social labels; and possible alternatives to present practice.

Hispanic Patterns of Test Scores.

It is obvious that, if a student does not understand the language in which a test is given, he will do a poor job of responding to the test items. If our goal is to arrive at a measure of "true" score in some trait, then we must ask whether by "true" we mean some long-run average of scores by this student, or whether we intend by "true" to suggest a genotype. For students from the majority culture, when the heritability of a measure is known, then we may infer the genotype by assuming that the phenotype (observed score) is correlated with this genotype as the square root of the heritability (if h^2 is the heritability estimate, then h is the correlation between genotype and observed score). This permits us, again with the majority student, to predict his genotype, within the error variance implied by h^2, in just the same way as we predict a long-run average of scores, knowing a single score.

With the Hispanic student, however, the discovered heritability does not hold, so far as English tests are concerned. Thus there is no way, from such a testing, to infer the student's genotype, or to make statements about his "true" score in this physiological sense of the word. If we mean his long-run average of such tests, however, or if we mean his score on other English-verbal tests not given to him, then we can reasonably infer his "true" score, given the known intercorrelations among these tests. In this sense, the English-language test is not "unfair" to him.

If we *do* wish to appraise the genotype, how may we do it? The usual procedure is to turn to other sorts of measures, requiring less or nothing of the biased material. Since at least World War I, there have been various "performance" or "non-verbal" measures available for supplementary information, either as part of the same IQ test, or as alternative indicants of mental function (in World War I, the verbal test was Army Alpha; the non-verbal, largely for illiterates, was Army Beta). Similarly, many current measures are available, and are widely used in special education. It is plausible, when wishing to know the effect of English language on intelligence test performance, that we compare verbal scores of an ethnic group with non-verbal scores, and we can express the discrepancy in an index used by Weyl (1969), 100 x NV/V.

When we look at some of the ethnic groups of most interest in the U.S., we find that, from the Coleman (1966) data, they can be ordered as follows:

Indian	110.9
Oriental	109.7

Mexican	107.7
Puerto Rican	120.0
White	101.6
Negro	95.6

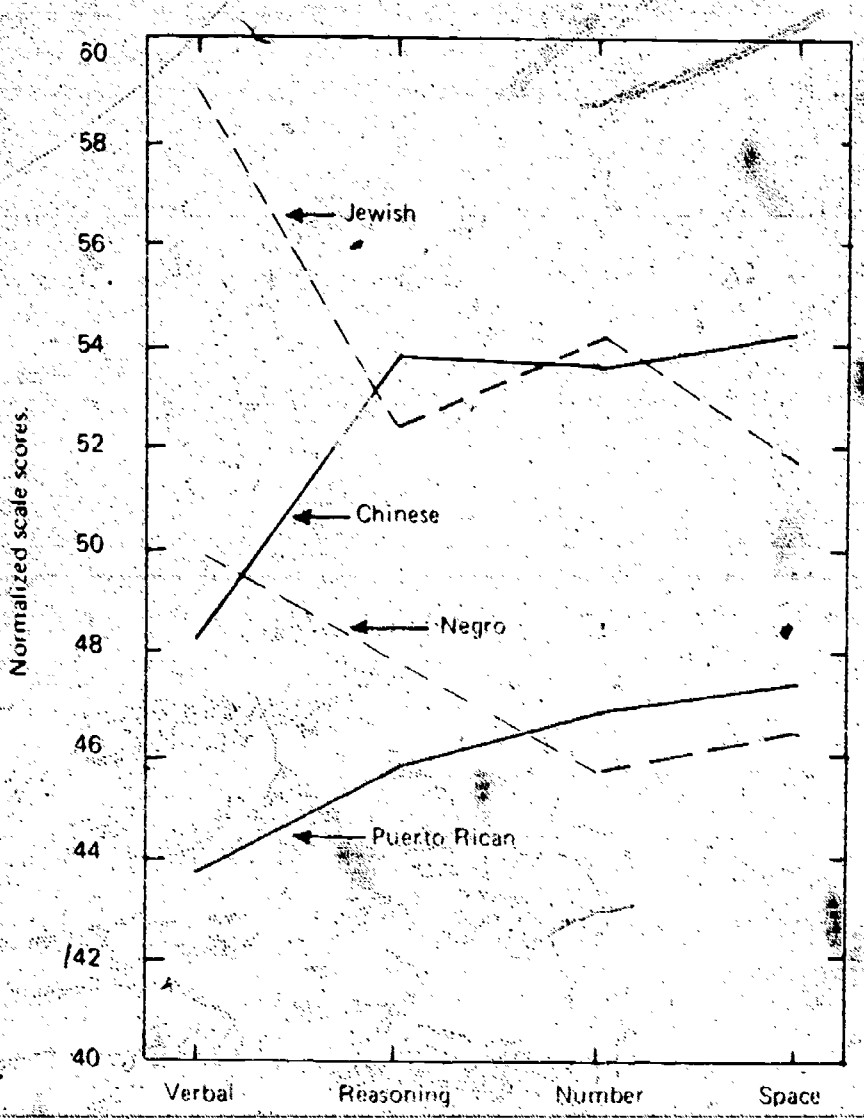
Let us assume for a moment that this ordering correctly suggests the relative language disadvantage of the groups. Then we see that Indians and Orientals are most hurt by English-language tests (compared with non-verbal alternatives or supplements), closely followed by Mexican-Americans, whereas Puerto Rican youngsters do little better on NV than on V tests, and Negroes seem to be, contrary to much popular and professional expression, actually favored by the verbal tests. (That the verbal and intellective skills are really quite independent is supported by the work of Furth [1964], who concluded that deaf children, though showing a large deficit in verbal testing, when such testing could be done at all, performed similarly to hearing children on non-verbal tests.)

To illustrate the complexity of the question, we examine the findings of Lesser, Fifer, and Clark (1965), for a large sample of New York City ethnic groups, as shown in Figure 19. Here we see the two foreign-language minorities, Chinese and Puerto Ricans, to be decidedly lower on the English-language IQ measures than on the less vocabulary-loaded traits of Reasoning, Number, and Space. The two English-language minorities, Jewish and Negro, show an advantage on the Verbal tests. But though the variance is definitely smaller on the less verbal tests, the groups are by no means equal in non-verbal performance. And evidence from other sources (Coleman *et al*, 1966; Loehlin *et al*., 1975, p.153; Jensen, 1971) shows similar group differences on both kinds of tests.

Clearly, then, if our goal is to be "fair" when discussing *genotype* for a broad category of "intelligence," we may not reasonably simply generate new norms for each ethnic minority and consider we have eliminated the problem. It does appear, however, that there is some definite English bias against Hispanics in assessing individual genotype for IQ. One suggested remedy might be to rely more on the NV tests for such appraisal. So far, evidence seems to suggest that the problem is not eliminated for Hispanics by simply administering the tests in Spanish, with Spanish-speaking test personnel. (In Hartford, Connecticut, these bilingual youngsters, mostly Puerto Ricans, did poorly on both English and Spanish versions of IQ tests.)

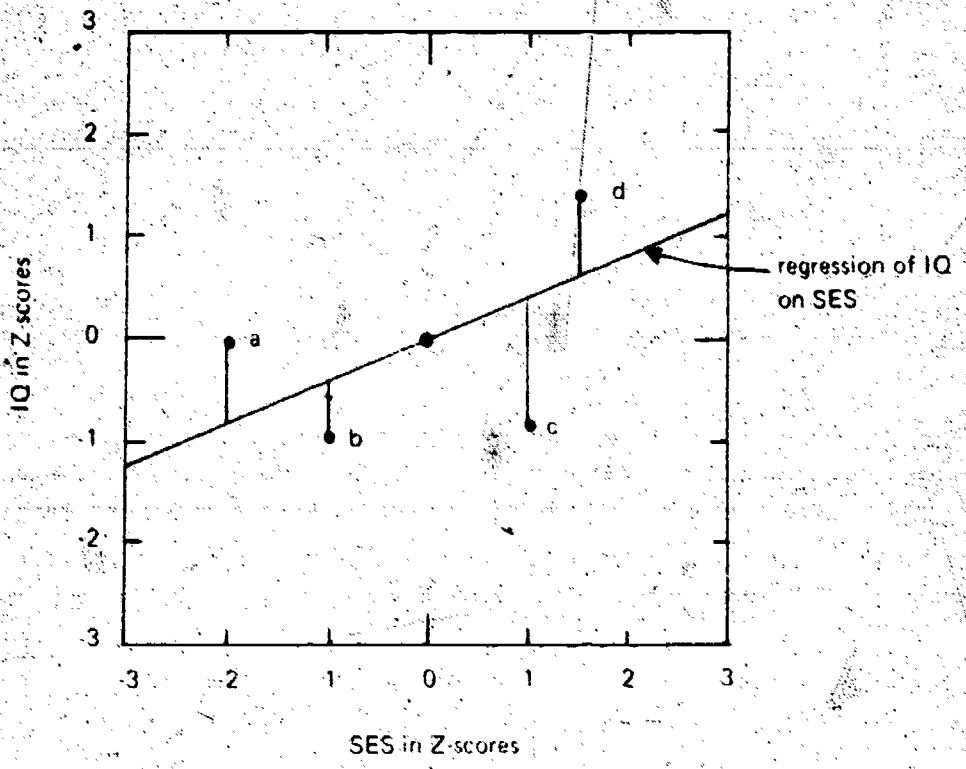
One strategy, sometimes recommended for establishing fair scores, is to make some adjustment for the SES of youngsters. In practice, given fairly reliable reports on income, parental occupation, and parental education, it would not be hard to estimate SES for each pupil and to make some adjustment for the known correlation between SES and IQ. The geometric interpretation of such an adjustment is as shown in Figure 20. In this figure, pupils *a,b,c,* and *d* would be reported in terms of their deviations above or below the regression line of IQ on

FIGURE 19
PATTERNS OF NORMALIZED ABILITY SCORES
FOR FOUR ETHNIC GROUPS



(From Lesser, Fifer, and Clark, 1965, reproduced in Loehlin, Linzey, and Spuhler, 1975.)

FIGURE 20
ADJUSTING IQ MEASURES FOR SES DIFFERENCES



In each case, the IQ of pupils a, b, c, and d would be reported in terms of the deviation from the regression line. Pupil *a*, for example, would be reported as having a superior IQ when so adjusted. (This is not recommended for reasons in text.)

SES. (Here the correlation is assumed to be about .40.) Pupil a's showing would be regarded as "superior" when so adjusted, and Pupil c's deficit is even more glaring after adjustment. When described in standard scores, the adjusted IQ would be calculated by the formula:

$$Z_{IQ(adj)} = Z_{IQ(obs)} - r_{SES\ IQ} Z_{SES}$$

where all terms are quite obvious.

Would such adjustment be "fair" when our concern is to express genotype? First, let us consider the question of such adjustment for majority (English-speaking, Caucasian) children. Heritability is well-established for such majority groups (for the treatment of the question in a best-selling psychology text, see Hilgard, Atkinson, & Atkinson, 1975, especially pp. 416-423). Also, there is of course a well-established relationship between parent IQ and parent occupation. The combination of these two widely acknowledged facts leads most scientists to the conclusion that lower SES children will have, on the average, lower IQ genotypes within the majority population (Herrnstein, 1973). As we have already seen, such partialling out of the variance for SES, then, throws out more than just environmental advantage; it also improperly removes variance attributable to genotype — that which we wished to measure.

When we consider the language-disadvantaged groups such as Hispanics and Chinese, the situation is not much better. The Chinese have done very well educationally and culturally in the U. S., and the SES adjustment does not seem necessary. As for the Hispanics, here adjustment for SES will not remove the central language problem, which is that in their homes English is seldom the only language (in only 19.7% according to Jensen, 1971). The reporting of non-verbal scores (clearly labeled as non-verbal) might be an appropriate step when the genotype is sought after; but it should be noted that the Hispanics will continue to trail behind the majority students (though for NV ahead of the Negroes). But not necessarily far behind: Jensen (1971) found the Mexicans in Berkeley to be mid-way on the Raven's test between the Whites and Negroes, despite having a "Home Index" (SES) as far below the Negroes as the Negroes were below the Whites. This suggests not *much* SES drag for the Mexican sample, but does not resolve the question of the "fairness" of inferring the genotype.

It is the tentative conclusion here, then, that other-language minorities may be reasonably described genotypically by use of the non-verbal tests, pending further advancement in either theory or data analysis.

I have gone to such length about the question of genotype of such youngsters for two major reasons: First, much of the writing attacking minority IQ testing centers around the question of "labeling" of the children as mentally retarded,

259

and the disproportionate number who are so classified. As we have seen, there will be far fewer Hispanics so labeled if the non-verbal tests are employed (though the percentage will still be slightly disproportionate, just as it will for the more successful minorities of Oriental, Jewish, or other groups). And in the language often used to attack such labeling, the question seems to be whether the children are "really" (i.e., genotypically) so retarded. A second reason is that the issue of genoticity of group differences has not been resolved, despite frequent allegations that it has, and should be recognized as underlying many of the court cases (e.g., *Larry P. v. Riles; Bakke v. University of California*) over the use of tests and the validity of affirmative action. But apart from these two reasons, much of the difficulty could be avoided by careful semantics concerning the tests and their use, as partly being described in the next section.

Diagnosing and Assigning Foreign-language Children

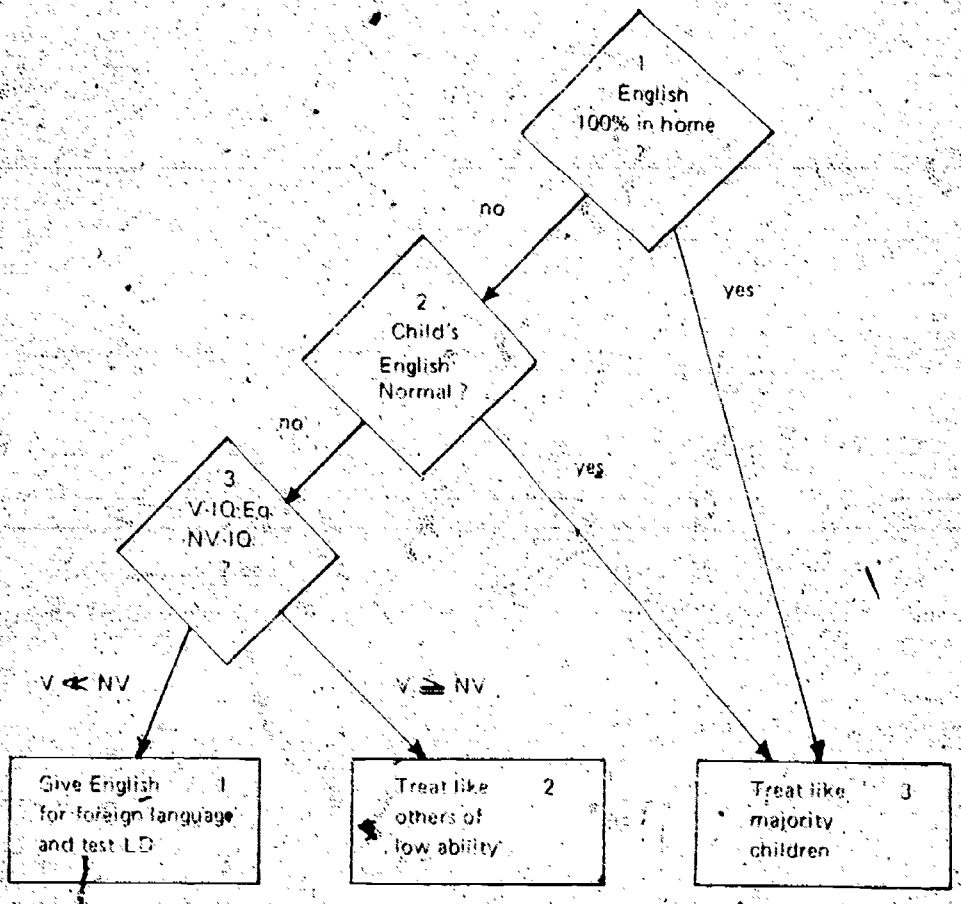
For purposes of individual guidance, there is probably only rarely any need to diagnose genotype. Therefore the term "IQ," loaded as it is with (often quite legitimate) implications of genotype estimation, is unnecessarily cumbersome and, when based on English tests for bilingual students, apparently biased as well. There are two cases which are at issue here:

1) Case 1 is whether the foreign-languge youngster is believed to be "culturally disadvantaged" because of this parental language. In a way, a disadvantage may be assumed and defined in terms of the percentage of time English is used in the home. Here it might well be that the status should be in terms of reading comprehension against the non-verbal tests of mental ability (whether *called* IQ, reasoning, "Raven's," or some other term would be, as we have seen, in part a matter of finding inoffensive usage). When the pupil is no longer much below his expected score in English reading comprehension, he may be incorporated into the ordinary programs throughout his school days.

2) Case 2 concerns whether the foreign-language child, no longer classified as "culturally disadvantaged," is in fact a victim of "learning diability" in one of the identified LD areas. In this second case, the proper comparison would seem to be *not* with his genotypic "mental ability" (by whatever name), but with his general, practical "verbal ability."

These two cases have been summarized in Figure 21, which is a suggested sequence of diagnosis for children of foreign-language background. The sequence consists of as many as three testing points, and three resulting action blocks. In the usual manner of flow-charts, the decision points are represented by diamonds, with alternative results; and the actions are shown by rectangles. In Decision 3, we note, there is testing of IQ (again, by whatever name) of both the Verbal and Nonverbal sort. Here, for purposes of decisions, it is quite appropriate that the Verbal testing be conducted in English (rather than in the child's "natural language"), since the assignment contemplated is for work in

FIGURE 21
SEQUENCE FOR FOREIGN-LANGUAGE STUDENTS



An emphasis on mainstreaming is here combined with separating language from other problems.

English as a second language. Or it could be a foreign-language IQ test, but, in the case of Spanish, it should not be expected that the performance of the bilingual child will be comparable with that of monolingual peers from the majority culture (Kahn, Note 3). In any case, this figure is presented for purposes of discussion and examination.

The diagnoses as social labels. As already noted, Verbal IQ scores do not have very sensible meaning when appled to minority-language children. Any findings from testing, then, should be clearly flagged as purely phenotypic (for measurement of "current language status"). Nonverbal IQ, on the other hand, can give qualified personnel help in deciding about genotype and, to some extent, about the probably outcomes of certain interventions. For example, if the NV-IQ is low, one will not expect the student to progress very rapidly in his mastery of English, and it may be that somewhat different techniques would be appropriate, possibly emphasizing basic rote features of English usage.

Here again, we have emphasized the *use* to which the scores are put, rather than their "label value." We should point out, however, that, however useful it may be to shift terms to something less demeaning, the value of the new terms is often quite short-lived. The term "moron", for example, is now seldom used, and "mentally retarded" seemed a relatively fresh and innocuous term. But terms describing low-status positions rapidly acquire the low status themselves: adolescent humor about "morons" has given way to adolescent humor about "tards." It is understandable that sociologists might wish to break out of this pattern, but linguistic psychologists should not be too optimistic. It is predictable that, in the fairly near future if not already, "learning disability" and "LD" will be terms of mild contempt, if not immediately of derision. And picking "completely neutral" labels, such as "X" or "Y", "bluebirds" or "redbirds", will not escape a certain heartlessness of the young, nor the anxieties of the parents, as experience in any tracked school system will ordinarily demonstrate.

Nor, it should be added, will "mainstreaming" avoid the labelling process of their peers and the communities. Relative comprehension will be exhibited. The extra attentions of the teacher will be noticed, and countless events will reinforce the impression of lower ability, and the test score, whether known or unknown, will probably play a relatively minor role in the way the child is regarded.

On the other hand, if the foreign-language child has normal abilities, or superior ones, these will also be made evident across time. The general attitude here, then, is that relabeling may be useful for a period, and may serve to change at least professional attitudes toward the handicapped student, whatever the source of handicap. But perhaps attention would be better directed toward understanding the labels and the conditions they denote, than toward semantic

262

reshuffling, or especially toward eliminating the preferred diagnostic instruments.

The Committee Seen as a Jury

Several parts of this report have emphasized the importance of personal judgment in making evalutions and placement decisions. We have already looked at the question of values in decision algorithms, and have seen that such values must depend on judgment. And we have considered as well the problem of agreement (or the lack of it) among judges, in the reliability of placement decisions. Yet one of the most difficult questions of judgment we have not touched on: what happens when we form individuals into a team, and then call upon that team for a single, unified action.

There is very little empirical research on such teams, in the field of Handicapped, and it is necessary to look to other fields for whatever useful evidence and theory we can abstract. One line of investigation which seems worthy of survey is that which has been carried out in jurisprudence, and which deals with the actions of juries in civil or criminal trials. This brief section could fit into a number of earlier sections of this report, but is unusual enough to be placed by itself, here at the end of Chapter III.

As we have seen, the Law requires the formation of pupil evaluation teams (PETs) to make decisions for the individual student. Such teams, at least whenever possible, are expected to consist of certain personnel: the pupil's classroom teacher, a specialist in the LD area considered, and someone competent in the measurement field, as well as other persons. When we analyze it, we realize that such a team will, then, act in a way analogous to that of a jury. While there is of course no "guilt" to be decided, the team will be estimating whether a pupil Johnny falls into the zone of the "disabled" and thus qualifies for such classification, or whether he does not. While "guilt" is often seen as a categorical variable, logically true or false, psychological placement is usually seen as establishing a cut-point for a continuous distribution. Yet a brief examination shows the two cases to be reasonably close together: many trials may hinge around the *degree* of guilt; and psychological placement does, after all, become reduced to a yes-no decision. Thus it may be instructive to turn briefly to a fairly substantial literature about decision-making in a jury.

The ideal data for study of jury trials do not exist, as noted by Gelfand and Solomon (1977). Such data would consist of independent replications of trials under varying arrangements of jury sizes and composition and other arrangements. As it is, they quote Nagel and Neef (1975) in their statement that jury decision-making is "a cross between flipping twelve independent unbalanced coins and bowling over twelve interacting bowling pins." The "independent

coin" model will be multiplicative in structuring the chance of particular outcomes, while the "interacting bowling pin" model will be additive in describing this chance. The true model would seem to lie somewhere in between.

But where? One way of examining the question is to see what is known about the jury action as compared with a single individual. As it happens, there is a great deal of evidence about jury verdicts compared with the hypothetical verdicts of the judges of the same trials. Kalven and Zeisel (1966) have collected data on thousands of such trials, and have summarized the concordance in Table 6.

TABLE 6
PATTERN OF JUDGE AND JURY DISAGREEMENT
FOR 3576 CASES BY PERCENTAGES

		Jury			
		Acquits	Convicts	Hangs	
Judge	Acquits	13.4	2.2	1.1	16.7
	Convicts	16.9	62.0	4.4	83.3
		30.3	64.2	5.5	100%

From Kalven & Zeisel (1966, p. 57), quoted in Gelfand & Solomon (1977, p. 297)

In these hypothetical verdicts, the judge is seen to be somewhat more inclined to convict than is the jury, even when one counts hung juries as being included with convictions. Of course, it would be improper to conclude, from the Table, that either judges or juries were "wrong" in the one-fifth of trials where they do not agree. There are good historical and philosophical reasons for both kinds of adjudication, just as there are obvious professional and political reasons for appointing the Pupil Evaluation Teams. And in both cases, the "truth" of the decision (whether the accused really is "guilty"; whether the pupil really is a victim of statutory "handicap") is not known. Rather, we have in the Table some evidence of the lack of perfect reliability in the data. In fact, when the Hangs are combined with the Convictions, and the *phi* coefficient is calculated (Hays, 1973, p. 743), we find an association of the judges and juries of .491. This way of considering a group decision, then, may be regarded as *external*, since it compares such decisions with those made outside the group, by others regarded, also, as having some validity.

Another way of regarding the process of the jury (or team) is *internal*, with

264

attention to what happens across time. In some hundreds of jury trials, investigators have collected information on the balloting of the jury; with particular attention to the first ballot and the final decision. The results, which may have some bearing on the functioning of PETs, are seen in Table 7.

TABLE 7
GUILTY VOTES ON FIRST BALLOT AND JURY DECISION
(N = 225 JURIES)

Final verdict	0	1-5	6	7-11	12	Total
Not guilty	100 %	91%	50%	5%	0%	32%
Guilty	0	2	50	86	100	62
Hung	0	7	0	9	0	6
	100	100	100	100	100	100

From Kalven & Zeisel (1966, p. 462)

All of the data from Table 7 were collected from 12-member juries, and are pretty strong in support of the persuasiveness of the majority. Where a minority entered the jury room believing the accused guilty, only 2% of the juries shifted to a final decision of guilty; whereas, when a majority of any size first believed the accused guilty, 86% of the trials were so concluded. On a three-person team, then, we might suppose that two of the members may have considerable influence over the third (although, as noted, the members of PETs each have their own presumed special perspective to bring to the question of placement).

From the sorts of evidence presented in the above tables, Gelfand and Solomon (1975, 1977) have designed a probabilitic model of some sophistication in order to analyze different sizes of juries. They were responding to a "claim" by the Supreme Court that a six-person jury may be expected to perform equivalently. They conclude, to the contrary, that in both the areas of representativeness of jury composition, and in the "quality" of decision-making, the larger jury will be more effective (1977, p. 311).

It is not suggested here that, for purposes of pupil assignment, any teams be considered as large as twelve, or even six. But repeatedly, we have pointed out that much more needs to be understood about the apparatus called upon to make these decisions, charged as they are with such cumbersome and legalistic and important responsibilities, and depending as they must on evidence which is

often quite shadowy and insubstantial. It is urged that BEH sponsor at least research, of both theoretical and applied nature, into how the decisions are, and should be, made in the school setting.

Public Accessibility to Records

In many areas of government today, there is a fundamental conflict between freedom of information, on the one hand, and the need for secrecy, on the other. In a sense, mental testing depends on surprise: Tests are designed to discover a *domain* of knowledge or ability, by checking out a *sample*. To the extent that the test sample is known in advance, its testing no longer permits an estimate of the domain. (This confusion between sample and domain leads to misunderstanding of behavior modification, as well.)

Even before P.L. 94-142, with its insistence on "public accessibility to records and information" and "parent involvement in plan development," there was trouble with the security of some standard tests, especially those administered individually for general mental ability. At times, items from the WISC have appeared in psychology textbooks (there read by average college students in this most popular major field). At times, items from WISC or other instruments have appeared in AP or UPI or other widely-distributed wire services. The "pick-a-fight" item for social intelligence was explicitly discussed on national television, in the CBS Report, "The IQ Myth." ("The CBS Myth" would have been a somewhat more accurate title.) What the effect of this sort of feedback is, especially among college-educated, middle-class families, remains very obscure; but it must be considerable.

Now, the new Law, with its requirements for parent involvement, public accessibility, "procedural safeguard guarantees," and several levels of appeal, makes probable a still greater difficulty in maintaining the item security of such instruments. In private communication, I have learned of remarkable abuses of security at the local level: One teacher told me that, when some pupils arrived for their testing for assignment to first grade, they may already have performed perhaps 50 "draw-a-man" exercises! With attendant instruction and feedback, such practice would wholly invalidate the norms for a pupil, yet such change would not imply *any* shift in school ability. A second teacher told me a similar story: Groups of up to 30 parents would gather to go over the specific items for the WISC or similar test. Again, pupils would expectably show a marked jump in overall score, by being practiced on the sample used, without any comparable change in the domain of knowledge or intelligence. How widespread are such practices is not known, but the provisions of the Law, and the important decisions depending on such testing, make still more probable parent anxieties and violations of test security on behalf of their youngsters.

A Parallel with College Admissions

But social decisions depending on testing are not new; and the threats to security of tests from public anxiety have been faced before – most notably in the well-established field of college admissions testing. Both the Scholastic Aptitude Tests (SAT) and the American College Testing Program (ACT) have designed very successful systems for admitting high-school seniors to colleges of mutual choice. And other programs successfully test applicants for admission to very selective, high-anxiety graduate studies such as law school (LSAT), medical school (MCAT), and doctoral programs (GRE). A nearly astonishing feature of such programs is the *absence* of major scandal involving release of test forms or group cooperation – astonishing, in view of the great aspirations which partly depend on high marks. It may be very instructive, then, to see just how these group tests, involving millions of testees every year, are managed and guarded.

Detailed procedures for such admissions testing are too lengthy to review here, but the outlines will be adequate for our purpose. Basically, these tests are founded on the idea of a *sample* of items from a specified domain. The tests are given only at specific times of the year, administered by outsiders who are paid by the testing companies, and each administration is surrounded by many safeguards. For example, the tests are to be kept in sealed packages until the actual test time. And once a test has been administered, it is regarded as no longer useful in just that form; most of its items will not appear again.

To support such a system, the test industry must constantly be in process of generating new items for the same target domain of abilities and knowledge. Item writers are always busy. And to keep the standards constant, the "meaning" of the scaled scores (typically 500 mean, 100 standard deviation), the industry has developed a very useful arrangement: In each testing period, there are some items "seeded" among the test items which will not, in fact, have any bearing on the applicant's score. These items are being tested out in a serious test situation for internal defects and for correlation with the already-tested items which are serving as the criterion of a student's score. Some of these seeded items will be abandoned, others rewritten, others used in a later, serious administration of the selection test.

Can we use these examples to improve our testing of the handicapped? I believe that, at the very least, it is a matter worth considerable study. After all, despite the traditions of individual tests such as the WISC, Stanford-Binet, and others, there is nothing sacred in the specific items employed. Let us consider vocabulary. It is typically the most heavily loaded sub-test in its correlation with *g* (general ability). In the S-B, the child is simply told to say what the words mean, and then the word is said: "orange." The tester tabulates the child's success. Even for the young and retarded, there are literally thousands of words

available for such testing, as youngsters rise in age and mental ability, the available vocabulary for testing becomes uncountably large. For vocabulary, then, it is possible to imagine changing the words frequently, using highly secure research programs to balance the meaning of the resulting IQs. Subsequent lists of vocabulary, then, would be similar to alternative *forms* of the same – all of them aimed at estimating the child's *true* vocabulary: the words understood in the domain.

For that matter, there is nothing sacred either about the *types* of items employed in the current tests. Not only are there many alternative possibilities within Block Design, there are also many alternative forms of items tapping essentially the same ability. Such types of items were, in the first place, a kind of cottage industry, invented and investigated by small numbers of individuals. But with such tests now determining "fair" classification into programs supported by billions in federal dollars alone, we seem to have left the stage of cottage industry; and individual testing, like group testing, should perhaps enter more of a big-industry structure.

There are questions, too, about just how "individual" such tests need to be. Within the limits of feasibility for the handicapped, programs should aim at useful scores through group testing, or through tests which, though administered one-on-one, do not require extensive background training in interpretation of the responses: if tests must be oral, they still may be multiple-choice, for example, or true-false. There are many adaptations which could be made for administration of tests by classroom teachers, for example, or teacher aides.

In other words, it seems that we can overcome this threat of the destruction of tests through the loss of security, by looking closely at the successful practices of college testing, and adapting these to the special needs of testing for the handicapped.

CHAPTER IV: RECOMMENDATIONS FOR IMPLEMENTATION OF THE LAW

In this chapter are grouped the "recommendations" which seem to follow from the analysis of the preceding chapters. It is reasonable to cluster these separately: it makes them easier for the practitioner to read and to reference. But also, it emphasizes the more tenuous quality of the recommendations, compared with the analysis itself. The statements in the analysis are set out with some confidence; they seem to this reviewer demonstrably "true." Actions, however, as we observed in Chapter I, depend on the balancing of many variables in an implicit decision analysis. Some of these variables include the subjective "values" attached to outcome, the differential weighting of various expected results. It is impossible for this reviewer (or any other outsider) to declare these values for the LEA. What follows, then, will be recommendations for what seem appropriate actions of the LEA, given a value system similar to my own:

1. *Determing ability.* It is recommended that each LEA establish a good data bank of test and other information about each suspected victim of handicap. It is recommended that "ability" be defined in terms of a *g* over the measures of mental ability, with the exception of the particular suspected area of handicap. The weighting of these tests into the *g* composite may be determined in a number of ways: publisher's recommendations (where a standard battery is concerned); publicly available factor analyses (in the published literature); or an a priori weighting (perhaps using the bentee token method) of the test scores. Such an apriori weighting may be done by a small number of the school system's researchers, psychologists, and special educators. Because this *g* will ordinarily be considerably more stable than a single IQ or similar test, the greater reliability will help in discriminating those who show severe discrepancy.

2. *Determining specific achievement.* Given the rather low reliability of many specific measures, a single test would not be enough evidence of treatable disability. It is recommended that the LEA establish a *construct* of the specific LD area by defining what is meant by the term, and then by weighting the various test scores which may be available for its diagnosis. Some of these will be subtests of already administered batteries; some may be inexpensive group instruments; others may require individual testing by a professional or para-professional. Again, the appropriate LEA personnel may use the token method to apportion the weighting of these measures, according to their reliability, and relevance to the LD construct. A single, composite score will then summarize the data for a particular child. This method will increase both the reliability and validity of the resulting LD measure. It will have the additional virtue of putting each individual test score in perspective, and emphasizing the variable nature of such scores.

3. *Determining discrepancy between ability and achievement.* As implied above, firm reliabilities should be secured for both ability and achievement composites. (These should really be considered across the total population, in the .90 area for test-retest reliability.) Once these are established, then it is reasonable to use discrepancy scores, based upon the formula $D = Z_{ach} - Z_{abil}$, where D is the difference or discrepancy, and Z represents the ususal standard score for achievement or ability (mean of zero, standard deviation of one). The reliability of D should always be reported with D, and the consequent band of confidence.

4. *Managing the analysis of discrepancy.* As explained in the text, the above requirements can require computer assistance. The LEA should either obtain such expertise locally, or should with other LEAs look to BEH or other appropriate agencies for a generalized set of computer programs. The details of such a program are beyond the scope of the present analysis.

5. *Reporting "true" scores.* As the earlier analysis made clear, the "true" scores are just as loaded with error, when one wishes to rank order students by disability, as are the observed scores. And the word "true" may create a misleading impression of accuracy, in the minds of counselors, parents, and teachers. If reliabilities are obtained of .90 for all reported measures, then there will not be much value in translating to a more moderate regressed score. Where it should appear valuable to state more moderate estimates, closer to the long-run average, then I'd favor the term "regressed score" or possibly the initials RS.

6. *Using judges to assess handicap.* This proposal is made seriously, but still tentatively, in view of the complexity of assessment of many different handicaps. It appears better, for the assessment of LD and possibly other problems, to make the final assessment mathematically, but *using* the judgmental ratings as part of the input. That is, let judges in the LEA appraise the student apart from any test scores, and preferably without knowing the test scores. Then take the judges' ratings, best expressed in numerical form, and include these ratings, with an appropriate weight, in the overall assessment. As pointed out in the text, subjective judgments are apt to be quite difficult to compare from one situation, problem, and student, to the next; yet they can furnish useful information. Thus, they should not be ignored, but neither should they constitute the final appraisal. They should be incorporated into this appraisal.

7. *Keeping judges independent.* At the assessment stage, it would be healthier if each judge evaluating a student did so independently, and recorded this, before sharing the rating with others. There is then plenty of opportunity to share different perspectives in the evaluation team. This recommendation stems from the need to develop understanding of the team process, and skill in it. Furthermore, these original ratings should be retained in the data system (possibly after removing identification) so that those LEA officials responsible for implementation might learn to understand such systems, and do applied research on them.

8. *Making decisions about placement.* What is described above is largely the process of measurement. Beyond measurement, is the necessary decision about placement. Ideally, this decision, too, should be made mathematically, through a weighted sum procedure. Ordinarily, in practice there will be some capacity within the system, and the decisions about "severity" will mirror this awareness of resource limitation. (It is unrealistic to assume that there is any cut-off independent of such resources.) Thus, for most LD situations, it will be a matter of rank-ordering the candidates for placement as well as one can, in light of *all* the variables (educational, psychometric, judgmental, familial, political) which may be estimated; and then selecting the most extreme students up to the capacity. What is urged is that the LEA, together with any help at the State and

Federal levels, design this procedure, or adapt an appropriate procedure which has been designed elsewhere.

9. *Allowing for appeals.* This is a thorny problem. As noted in the text, if the appeal is against the assignment to a special-education treatment, then the appeal is likely to win, in the case of a truly independent judgment. On the other hand, if the appeal is against an assignment to a regular classroom, then the appeal is likely to lose, under these same conditions of independence. The tendency of having appeals, then, is to exclude more students from the treatments reserved for the extremes. Since I strongly favor an ultimately mathematical system which uses the subjective as input, I would suggest that the appeal process not aspire to independence (it would be virtually impossible), but only to a more informed alteration of *certain* of the input variables, then recalculate the decision with these new values. This will make possible: 1) gathering of new information; 2) reconsideration of the subjective variables, such as familial support; and 3) recalculation of the decision. But it will prevent the Appeal level serving rather blindly as a reversal of inclusion or, equally blindly, as supporting the earlier decision to stabilize the system.

10. *Maintaining records of PET actions.* All must be aware that compliance with the Handicapped Law will be awesomely expensive for all concerned. A relatively trivial expense will be aggregating a bank of data which will permit applied and basic research into the local situation and the more general functioning of the teams. At least 7% of the cost of testing, diagnosing, assessing, and assigning should go into maintaining such records. The design of the system, however, should not be the responsibility of the LEA; the forms should be standardized, and quantified throughout, for the pooling of data across the LEAs and across the SEAs. The determination of desired variables and their representation on the forms should be done by centralized agencies, in close partnership with a sample of representative SEAs and LEAs.

11. *Maintaing records of treatments and outcomes.* The same arguments support a system of data following the Handicapped across the years, including treatments, post-treatment testing and appraisal, and some follow-up for selected samples of such students. Again, it is recommended that at least 7% of the total cost of the LEA Handicapped Program be used for such a system. (Note: these 7% recommendations are not intended to include ordinary case records and guidance; rather, these are intended for the additional, controlled, *researchable* records useful for evaluating the treatments and diagnostic procedures and suggesting ways in which the laws, guidelines, and actions may be improved. It should be recognized now just how vastly ignorant we are of the utility of these mandated arrangements. It is predictable that the systems will be modified, say, 30% in the next few years; it is only cost-effective, therefore, to make a major allocation of resources to monitoring the early approximations.)

271

12. *Collecting cultural data.* Since the Handicapped Law insists that assessment should be free of ethnic and cultural bias, it is mandatory that ethnic and cultural data be collected for appraising any such bias. Attached to each researchable record, then, will be standardized information about ethnic membership, languages in the home, parental education and occupation, and some estimate of range of income (the standard SES variables). Again, these should be collected in a standardized way, for comparisons across states and localities. Designing the standard forms, then, should be a centralized responsibility.

13. *Collecting sibling information.* Questions of "bias" for the PEP inescapably involve family variables, since the family, one way and another, typically accounts for much more of the variation in ability and school achievement than any within the control of the external society. For purposes of understanding the sources of the handicap, then, and the students' responsiveness to treatment, it is very important to take advantage of the presence of siblings, especially of twins. Within each hundred children in a school system, at a given age-level, there will be at least one pair of twins, who have shared almost from conception the same general environment, and to a varying degree the same inherited constitution. These twins are easily identified in school records through having the same family names, the same address, and the same birth-date. If we are to understand handicap and what we can do socially to overcome it, it is important to identify the sources which are most under environmental control. It is recommended, therefore, that the standard forms designed for record-keeping of the Handicapped routinely raise the question about the existence of a twin. When located, test information should be collected on the twin as well, and a simple form used to estimate kinship (whether identical or fraternal). If the twin is also being evaluated for handicap, then a flag should be put in the standardized records, typing the two records together for research purposes. If the twin is not being so evaluated, then the standard test information, ordinarily collected by the school system, should be simply incorporated into the cultural background of the handicapped twin. Such twin information may prove, for that sample of students on whom it is collected, to be the richest source of understanding for the later adjustment of the total system.

14. *Adjusting assessment for Negro students.* In most researches, White and Black students share the same regression lines when academic attainment is compared to general cognitive testing. Therefore, it is *not* recommended that special norms be used for Black students; to the contrary, such special norms would probably diminish the utility of the testing and assessment strategy. When it comes to placement, a color-blind policy, using the same formulation, is recommended. However, as noted above, there is a (limited) role of subjective appraisal in such a weighted-sum procedure, and to the extent that a Black youngster is in fact a higher achiever than his tests would indicate, there would be adjustment to this fact (as there would be for a White youngster, as well). As

272

noted in the text (Fig. 18), this would probably produce a somewhat larger proportion of LDs who are Black, for the statistical reasons outlined.

15. *Adjusting assessment for different-language students.* A recommended sequence of diagnosis for foreign-language students is outlined in Figure 21, elsewhere in this text, and will not be repeated here. I believe that all assessment and placement of foreign-language youngsters, whether Latins or from the groups which are more successful academically, should be aimed at the maximum integration of the children into English-speaking classrooms. In general, rather than use the clearly inappropriate verbal IQ tests to measure general ability, I would recommend using the non-verbal ability measures. From available evidence, this will still produce a larger proportion of Latin students in classes of slower students, but not nearly the same number as would be so placed with English (or Spanish) verbal tests. It is *not* recommended that separate norms be established for the Spanish-speaking groups, and then IQs artificially inflated to a normal distribution, without any matching of these spurious "abilities" with corresponding achievements. Such statistical manipulation will only lead to confusion and false expectation on the part of teachers, parents, and the students themselves. However, in most cases the question of IQ is not central. A more urgent concern is the speedy training in English as a second language required to mainstream Latin youngsters into the regular classrooms, where they will be soaked in the majority language and culture.

16. *Protecting the assessment tests.* As noted in the text, many items of many tests are being spoiled for use by the destruction of their security. The LEA personnel should understand that they are ethically required to protect test security of measures in use for assessment under the Handicapped Law. Parents are welcome to know the *types* of tests used, but not necessarily the particular tests, and most surely not the specific items employed in the past or to be used in the future. The fairness of the Handicapped system depends in part on such tests coming as fresh to all students on whom they are to be employed.

These are only some of the principal recommendations which seem to follow from the preceding technical analysis of selected features of the Handicapped Law and guidelines. Unfortunately these recommendations, like the guidelines themselves, will not appear to be immediately implementable in the schools, without considerably more detailed study and comment. And no doubt, some will seem infeasible for political or other reasons explicitly excluded from this psychometric appraisal.

What is important to remember is that the present guidelines, and their interpretations in practice, are surely not carved in stone. All of us who work in education need to bring our best analysis to bear on them, and on their effects. Especially, we need to develop information feedback which will permit us to monitor them and their effects, both beneficial and otherwise, across the decades ahead.

273

REFERENCES

Atkinson, R. C. Ingredients for a theory of instruction. *American Psychologist,* 1972, *27,* 921-931.

Bereiter, C. Genetics and educability: Educational implications of the Jensen debate. In J. Hellmuth (Ed.), *Disadvantaged Child.* Vol. 3, *Compensatory education: A national debate.* New York: Brunner-Mazel, 1970. Pp. 279-299.

Blalock, H. M., Jr. (Ed.) *Causal models in the social sciences.* Chicago: Aldine, 1971.

Block, N. J., & Dworkin, G. (Eds.) *The IQ controversey: Critical readings.* New York: Pantheon, 1976.

Cancro, R. (Ed.) *Intelligence: Genetic and environmental influences.* New York: Grune & Stratton, 1971.

Cleary, T. A., Humphreys, L. G., Kendrick, S. A., & Wesman, A. Educational uses of tests with disadvantaged students. *American Psychologist,* 1975, *30,*

Coleman, J. S., *et al. Equality of educational opportunity.* Washington, D. C.: U. S. Office of Education, 1966.

Cooley, W. W. Techniques for considering multiple measurements. Chap. 16 in R. L. Thorndike (Ed.), *Educational measurement,* 2nd ed. Washington, D. C.: American Council on Education, 1971. Pp. 601-624.

Cronbach, L. J., & Gleser, G. *Psychological tests and personnel decisions,* 2nd ed. Urbana: Univ. of Illinois Press, 1965.

Falconer, D. S. *Introduction to quantitative genetics.* New York: Ronald Press, 1960.

Findler, N. V., Pfaltz, J. L., & Bernstein, H. J. *Four high-level extensions of FORTRAN IV.* New York: Spartan, 1972.

Furth, H. G. Research with the deaf: Implications for language and cognition. *Psychological Bulletin,* 1964, *62,* 145-164.

Gelfand, A., & Solomon, H. Analyzing the decision-making process of the American jury. *Journal of the American Statistical Association,* 1975, *70,* 305-310.

Gelfand, A. E., & Solomon, H. Considerations in building jury behavior models and in comparing jury schemes: An argument in favor of 12-member juries. *Jurimetrics Journal*, 1977, *17*, 292-313.

Hays, W. L. *Statistics for the social sciences*, 2nd ed. New York: Holt, Rinehart and Winston, 1973.

Hebert, J. P. *Race et intelligence*. Paris: Copernic, 1977.

Herrnstein, R. J. *IQ in the meritocracy*. Boston: Little, Brown, 1973.

Hilgard, E. R., Atkinson, R. C., & Atkinson, R. L. *Introduction to Psychology*, 6th ed. New York: Harcourt Brace Jovanovich, 1975.

Hills, J. R. Use of measurement in selection and placement. Chap. 19 in R. L. Thorndike (Ed.), *Educational measurement*, 2nd ed. Washington, D. C.: American Council on Education, 1971. Pp. 680-732.

Jensen, A. R. Do schools cheat minority children? *Educational Research*, 1971, *14*, 3-28.

Jensen, A. R. *Educability and group differences*. New York: Harper & Row, 1973.

Kalven, H., & Zeisel, H. *The American jury*. Boston: Little, Brown, 1966.

Keeney, R. L., & Raiffa, H. *Decision with multiple objectives: Preferences and value tradeoffs*. New York: Wiley, 1976.

Lesser, G. S., Fifer, G., & Clark, D. H. Mental abilities of children from different social-class and cultural groups. *Monographs of the Society for Research in Child Development*, 1965, *30*, No. 4.

Loehlin, J. C., Lindzey, G., & Spuhler, J. N. *Race differences in intelligence*. San Francisco: Freeman, 1975.

Loehlin, J. C., & Nichols, R. C. *Heredity, environment, and personality: A study of 850 sets of twins*. Austin: Univ. of Texas Press, 1976.

Martin, N. G. The inheritance of scholastic abilities in a sample of twins. II. Genetical analysis of examination results. *Annals of Human Genetics (London)*, 1975, *39*, 219-229.

Mather, K., & Jinks, J. L. *Biometrical genetics*, 2nd ed. Ithaca, N. Y.: Cornell University Press, 1971.

McClearn, G. E., & DeFries, J. C. *Introduction to behavioral genetics.* San Francisco: Freeman, 1973.

Mercer, J. R. Sociocultural factors in the educational evaluation of Black and Chicano children. In Select Committee on Equal Educational Opportunity, *Environment, intelligence, and scholastic achievement.* A compilation of testimony. Washington, D. C.: U. S. Government Printing Office, 1972. pp. 438-449.

Mercer, J. R. *Labelling the mentally retarded.* Berkeley: Univ. of California Press, 1977.

Millman, J. Passing scores and test lengths for domain-referenced tests. *Review of Educational Research*, 1973, *43*, 205-216.

Millman, J. Sampling plans for domain-referenced tests. *Educational Technology*, 1974, *14*, 17-21.

Minsky, M., & Papert, S. *Perceptrons: An introduction to computational geometry.* Cambridge: M. I. T. Press, 1969.

Nagel, S., & Neef, M. Using deductive modeling to determine an optimum jury size and fraction required to convict. Paper presented to Annual Meeting of American Society for Public Administration, 1975.

Nichols, R. C. Policy implications of the IQ controversy. In L. S. Shulman (Ed.), *Review of Research in Education*, 1978.

Page, E. B. How we *all* failed in performance contracting. *Educational Psychologist*, 1972, *9*, 40-42. Also in *Phi Delta Kappan*, 1972a, *54*, 115-117.

Page, E. B. Seeking a measure of general educational advancement: The bentee. *Journal of Educational Measurement*, 1972b, 9, 33-43.

Page, E. B. Effects of higher education: Outcomes, values, or benefits. In Solomon, L. C., & Taubman, P. J. (Eds.) *Does college matter? Some evidence on the impacts of higher education.* New York: Academic, 1973. Pp. 159-172.

Page, E. B. Problems and perspectives in measuring vocational maturity. In D. E. Super (Ed.), *Measuring vocational maturity for counseling and evaluation.* Monograph of the National Vocational Guidance Association, 1974a. Pp. 68-79.

Page, E. B. 'Top-down' trees of educational values. *Educational and Psychological Measurement*, 1974b, *34*, 573-584.

Page, E. B. Evaluating 'evaluation.' *Review of Education*, 1975, *1*, 133-140.

Page, E. B. A historical step beyond Terman. In D. P. Keating (Ed.), *Intellectual talent: Research and development*. Baltimore: Johns Hopkins Press, 1976a. Pp. 295-307.

Page, E. B. The optimization of educational values in Navy curriculum design. Presented at Annual Meeting of the American Statistical Association, August 26, 1976, Boston. *Proceedings of the Social Statistics Section of the ASA*, 1976b, Part II: 655-659.

Page, E. B., & Breen, T. F. III. Educational values for measurement technology: Some theory and data. In W. E. Coffman (Ed.), *Frontiers in Educational Measurement and Information Processing*. Boston: Houghton-Mifflin, 1974. Pp. 13-30.

Page, E. B., & Grandon, G. M. Family configuration and mental ability: Confluence vs. admixture theories. Manuscript submitted for publication, July, 1977. Pp. 43.

Page, E. B., Jarjoura, D., & Konopka, C. Curriculum design through operations research. *American Educational Research Journal*, 1976, *13*, 31-49.

Raiffa, H. *Decision analysis: Introductory lectures on choices under uncertainty*. Reading, Mass.: Addison-Wesley, 1968.

Salvia, J., & Clark, J. Use of deficit to identify the learning disabled. *Exceptional Children*, 1973, *39*, 305-308.

Scarr-Salapatek, S. Race, social class, and IQ. *Science*, 1971, *174*, 1285-1295.

Shuey, A. M. *The testing of Negro intelligence*, 2nd ed. New York: Social Science Press, 1966.

Slagle, J. R. *Artificial intelligence: The heuristic programming approach*. New York: McGraw-Hill, 1971.

Stanley, J. C. Predicting College success for the educationally disadvantaged. *Science*, 1971a, *171*, 640-647.

Stanley, J. C. Reliability. Chap. 13 in R. L. Thorndike (Ed.), *Educational measurement*, 2nd ed. Washington, D.C.: American Council on Education, 1971b. Pp. 356-442.

Swaminathan, H., Hambleton, R. K., & Algina, J. A Bayesian decision-theoretic procedure for use with criterion-referenced tests. *Journal of Educational Measurement*, 1975, *12*, 87-98.

Tillett, P. I. An operations research approach to the assignment of teachers to courses. *Socio-Economic Planning Sciences*, 1975, *9*, 101-104.

Trueman, R. E. *An introduction to quantitative methods for decision making*. New York: Holt, Rinehart & Winston, 1974.

VanDusseldorp, R. A., Richardson, D. W., & Foley, W. J. *Educational decision-making through operations research*. Boston: Allyn & Bacon, 1971.

Wagner, H. B. *Principles of operations research with applications to managerial decisions*. Englewood Cliffs, N. J.: Prentice-Hall, 1969.

Weyl, N. Some comparative performance indexes of American ethnic minorities. Mankind Quarterly, 1969, *9*, 106-128.

Wilcox, J. W. *A method for measuring decision assumptions*. Cambridge: M. I. T. Press, 1972.

Wilson, A. B. Educational consequences of segregation in a California community. In *Racial Isolation in the Public Schools*, App., Vol. II, Pp. 185. Washington: U. S. Commission on Civil Rights, 1967.

Ysseldyke, J. E. Current issues in the assessment of learning disabled children and some proposed approaches to appropriate use of assessment information. Paper presented at the BEH Conference on Assessment in Learning Disabilities, Atlanta, May, 1977.

Ysseldyke, J. E., & Salvia, J. Diagnostic prescriptive teaching: Two models. *Exceptional Children*, 1974, *41*, 181-186.

PART C

The View from the Panel

INTRODUCTION

The 2-day panel meeting provided an opportunity to bring together a small but diverse group of educators to react to both the study and the protection in evaluation procedures position papers. The group included representatives from state and local education agencies, university departments of special education, and the Bureau of Education for the Handicapped (BEH). Following initial BEH presentations by Dr. Linda Morra and Dr. Mary Kennedy, which set the general context for the study, authors presented summaries of their papers and responded to questions and comments. During the afternoon, panel members discussed various issues related to the study and specific papers. On the second day, small groups were formed to continue discussion of issues and develop recommendations. A general session followed to share the results of the small group sessions. An issue-by-issue summary of the panel discussion and a summary of the small group recommendations are presented in the next sections.

THE ISSUES

Panelists did not view assessment as a single procedure forming one step of the special education decision-making process. Instead child assessment information was viewed as critical to the making of each decision in the process. Thus, the need for protection in evaluation procedures throughout the special education decision-making process was evident to panelists. The central theme of the panel became the implementation of protection in evaluation procedures to prevent abuse in decision-making. Panelists identified four major questions to be addressed in the decision-making process, each of which should be based on child assessment information. Panelists discussed needed protection in evaluation procedures related to the assessment undertaken to obtain information to respond to each question. The relationship between the decisions and protection in evaluation activities identified by the panel are shown as follows:

Decision Requiring Assessment Information	*Protection in Evaluation Procedures*
1. Should the child be referred for a psychoeducational evaluation?	a. Conduct routine screening of children. b. Analyze teacher referral patterns. c. Analyze previous educational interventions attempted
2. Is the child eligible to receive special education and related services?	d. Examine the adequacy of the evaluation undertaken to determine eligibility

Decision Requiring Assessment Information	*Protection in Evaluation Procedures*
3. What is the child's handicapping condition?	e. Consider the utility of this information.
4. What should be the child's education program?	f. Determine the adequacy of the evaluation undertaken to make educational interventions.
	g. Review the continued appropriateness of the child's special education program

The presentation of the issues identified by panelists begins with a general discussion of problems in implementation of the PL 94-142 protection in evaluation procedures. It then proceeds with discussions of other protection in evaluation procedures issues, as outlined above, which are related to the special education decisions. The section ends with a discussion of the utility of self-study guides for address of the protection in evaluation procedures issues.

Problems in Implementation

Central to the discussion of PL 94-142 protection in evaluation procedures (PEP) was the issue of implementation at the local school district level. The regulation requiring that tests be validated for the specific purpose for which they are used was thought by panelists to be particularly problematic. As stated by one panelist: "I conduct workshops with school psychologists and I can ask how many of you use standardized tests? And all hands go up. Then I ask on what groups were those tests standardized, and no hands go up. People have not looked at the test standardization and they have not thought carefully about the extent to which the individual they assess has a cultural background comparable to those on whom the tests were standardized." A related problem was inappropriate modification of tests for handicapped individuals. One example of this practice would be the administration of an intelligence test, standardized on a hearing population, to a deaf child through manual sign language. Another area of concern expressed by panelists was the use of assessment devices with reliability coefficients as low as .12 in the making of decisions which significantly effect the lives of children. Yet another problem identified was use of intelligence test subtest scores to make specific educational prescriptions.

The basic problem was identified of implementing current assessment standards in practice. There was general consensus among the group that principles of

assessment now exist, such as those embodied in the American Psychological Association standards. If these standards were applied in practice, substantial progress would be made towards elimination of some of the current discriminatory practices in assessment.

The Referral Decision

A major theme in the large group discussion was the need for protection in evaluation procedures at the point of consideration of children for referral for an individual evaluation. Errors in the direction of both under-referrals and over-referrals were felt to occur. One panelist illustrated the possible problems in the referral process: "Somebody in the public schools, typically a teacher, decides a kid has a problem. What are the observations which lead to that conclusion? What are the potential sources of bias in these initial observations and judgements?"

Those panelists concerned with large increases in the numbers of children referred for evaluation associated much of the difficulty with lack of tolerance for deviant behavior in the regular classroom. A general recommendation of these panelists was for some kind of validity check on referrals. As stated by one panelist: "A lot of assessments are not necessary. The first consideration should be making minor adjustments in a kid's program." For example, one panelist stated that some school districts simply ask the teacher upon referral of a child to list five specific approaches he/she took in trying to resolve the problem concerning the child. Another school district described by a panelist requires teachers to sign their name to a statement on the referral form to the effect that the child is not benefiting sufficiently from regular classroom instruction. Both strategies were viewed as reducing the number of referrals made by teachers.

From the perspective of the Bureau of Education for the Handicapped, the more serious problem stems from under-referrals and the possible lack of provision of special education and related services to those children who require them. Those panelists concerned with under-referrals felt strategies were needed to encourage teacher referrals. One panelist provided the following anecdote to illustrate this point: "We conducted an experiment where we went into three school districts and tested all the first and third graders. We were worried about teachers referring too many kids. When we cross-checked and compared the numbers of students teachers had referred [with the number we felt should have been referred] we were shocked at how low the number of teacher referrals was. It was nowhere close to what we [school] psychologists would have referred—even [restricting] our referrals, to the most obvious and glaring cases." Research results were also cited which indicate that screening efforts and teacher referrals typically identify two different groups of children for evaluation. This finding was viewed as providing support for the use of routine screening procedures.

Additionally, studies were described in which no differences were found on test after test between children having difficulty in regular class, some of whom were labeled as handicapped and some of whom did not carry the label. It was found, however, that children carrying the handicapped label displayed behaviors which teachers found particularly troublesome. Panelists felt that typically withdrawn children would be overlooked, while acting-out children would be referred. One suggestion offered by a panelist was that teacher referral patterns be examined for possible biases such as under-referrals of certain types of children.

There was some agreement among panelists that regardless of whether the problem was perceived as over or under-referrals, there was a need for in-service training of regular classroom teachers to reduce bias in the referral decision. One suggestion was for in-service training which would focus on developmental psychology. Another suggestion involved a socio-psychological approach in which people would learn to recognize their own biases in making referral decisions. As discussed, there was also general support by panelists for the implementation of routine screening procedures.

Adequacy of Child Evaluations for Eligibility and Programming Decisions

A related concern was the adequacy of the evaluations being conducted either to determine a child's elegibility for special education services or to determine a child's special education program. A specific area of concern was the time-consuming nature of thorough interdisciplinary assessments of children. The panelists agreed that ideally, teachers, school psychologists, audiologists, nurses, and other professionals should be involved in the diagnostic evaluations. In addition, assessments should be multifactored with measures of adaptive behavior, observation of the child in natural settings, medical information, self-reports, and the like. The time-consuming nature of such procedures, however, coupled with a high number of referrals, desire to process the referrals quickly, and not enough diagnosticians for the number of referrals were viewed as often resulting in less than ideal implementation of assessment procedures.

One possible strategy suggested for resolving these problems was to establish case load ratios for the various categories of diagnosticians. While each diagnostician is likely to have other responsibilities in addition to assessment, a clear time block would be defined for assessment activities. The number of cases which could be handled in the time available could be determined, and projections of additional staff needed would also be easily determined.

The Classification Decision

A major issue in the panel discussion was the use of categorical labels to discriminate among children in need of special education and related services. While panelists recognized that labels are necessary for PL 94-142 funding purposes, a major question was whether, in fact, the labels lead to specific treatments or program interventions. One panelist expressed the question as: "Do we have different ways of treating a child who is labeled mentally retarded as opposed to learning disabled?" It was not clear to panelists that categorical special education programs are always differentiated in this respect. However, discussion of alternative approaches to labeling led to other unresolvable problems.

For example, one panelist suggested that children be described only in terms of the services they require based on the individualized education program. A consequence found with this approach, however, has been that a large number of children are identified as needing special services who were not eligible for special education and related services under the system of categorical definitions of handicapping conditions. The problem, as posed by one panelist, is that state aid requirements for the handicapped are likely to double. While other panelists felt that all children in need of specific services should be provided those services, they doubted that the problems of labeling would be avoided. If, as stated by a panelist, a child is in remedial reading program 4, the child will soon be known as a "program 4" child.

At least one panelist felt that research was needed on educational interventions or treatments and the achievement of different groups of students. Specifically, this panelist recommended a longitudinal study of teacher-student interactions and their effects on low, average, and high achieving groups of students. Another panelist, however, felt that aptitude-treatment interaction research has shown little promise to date.

The panel did not resolve the issue of labeling and come to any recommendation. For now, as stated by one panelist, "we are required by PL 94-142 to document the label – we are accountable diagnostically."

Assessment for Educational Programming Decisions

Panelists stressed the need for clearer articulation of the linkages between assessment and program intervention. Individual evaluations were viewed as traditionally conducted to determine a child's eligibility for special education and related services or, in more general terms, for classification purposes. While need was recognized for standard assessments which identify children who

require special educational intervention, panelists emphasized the need for assessments which also determine what specific interventions in the child's educational program should be made.

There was general consensus that different assessment data were needed for eligibility decisions than for programming decisions. As stated by one panelist: "A factor such as 'g' will do an excellent job of predicting who will and will not succeed in school. It won't do a very good job of showing how specifically to teach an individual child." Norm-referenced tests were viewed as appropriate assessment instruments for eligiblity decisions, and criterion-referenced tests as appropriate assessment instruments for programming decisions. The panel generally agreed that assessments should include the multiple types of data and, following the PL 94-142 regulations, include those tests tailored to assess specific areas for educational need and not merely those designed to produce a single general intelligence quotient.

One panelist suggested that linkages among assessment, program intervention, and program review activities would be strengthened if the evaluation team consisted of the same persons who develop the child's individualized education program (IEP). The team was described as including, at a minimum, a teacher, a special education specialist with expertise in individual educational programming, a school psychologist, and parent of the child. Another panelist warned, however, that more research and test development was needed before faith could be placed in relationships between test performance and educational programming. This panelist cited a commonly used diagnostic instrument. One could work with the child and raise the child's score on the test, but this does not mean that the child reads any better. While there was agreement that more research was needed in this area, other panelists stated that with the use of criterion-referenced tests and the systematic application of the principles of learning, "we can move kids forward instructionally."

Self-Study Guides

There was agreement among panelists that the development of procedures to evaluate implementation of the PEP procedures was a useful endeavor. There were, however, different views concerning evaluation methods and content. Several panelists recommended the development of a checklist or guide which could be used by school districts on a self-study basis. It was pointed out by one panelist that many school districts still essentially use the IQ test for decisions concerning eligibility and programming and the guide or checklist would be helpful as a means of improving practice. Another panelist was of the opinion that a guide would be too limited in use if it was focused towards the below average or average district in terms of PEP implementation. Another suggestion offered a full evaluation procedure. A district would take the names of a random

sample of handicapped children and collect information on what had happened to those children. Information collected would include referral forms, the assessment and diagnosis, the IEP, and performance results. Among the assessment information that might be checked, for example, would be the validity of the instruments for the purposes they were used. A final point made by one panelist was that if change is the goal, school districts should be involved in the process of developing the guide.

RECOMMENDATIONS

All three of the panel subgroups recommended the development of either technical assistance guides or models which would offer implementation and/or evaluation strategies to local school districts concerned with the protection in evaluation procedures provisions of PL 94-142. The groups differed, however, in their descriptions of the focus of guides or models. Group I recommended the development of a conceptual framework which would outline a total child planning and programming system. For each stage, issues, critical events, self-evaluation strategies, and possible corrective actions would be identified. Each state and school district, however, would have to adapt the guide to meet its own needs. Group II, which had many recommendations similar to those of the first group, recommended a self-study guide which would include sections on compliance, monitoring, and evaluation of the PEP provisions. Rationales, issues, standards, good-practice examples, and criteria would be identified and discussed. Group III focused on developing a model for fair decision-making procedures concerning the educational placement of children. The model, which utilizes a cascade system, closely relates assessment to programming issues.

Group I

Panelists in Group I took the position that assessment activity should not be viewed as a separate component in the child planning and programming process. They recommended that assessment activities be examined in relation to each stage in the child identification and planning process. Stages were identified as (1) prescreening, (2) screening, (3) referral, (4) individual evaluation, and (5) IEP development and implementation. The last stage includes placement. Panelists recommended the development of a technical assistance document which would detail assessment activities in each stage.

The group provided an example of a conceptual framework for such a technical assistance document. For each of the above stages in the child identification and planning process, procedural issues, critical events, self-evaluation strategies, and examples of possible corrective actions would be described. These four topics, where appropriate, would be addressed separately for the school district and

individual case levels. The group also recommended that in the document distinction be made between PL 94-142 compliance requirements and implementation procedures which are exemplary and not required.

The group illustrated how the conceptual matrix might be operationalized. The prescreening stage, for example, refers to record review of children new to the school district. An evaluation strategy at the school district level might be to check the adequacy of information that can be obtained from kindergarten registration forms for identifying known or suspected disabilities such as a hearing or vision problem. The section on corrective actions in this case might include examples of registration forms currently used by school districts which are useful for identifying children with potential special education needs.

At a case level, an evaluation strategy might randomly select a number of the permanent record files of transfer students to determine if a review of the child's former school records was conducted when the child entered the system. In the referral stage, to provide another example, procedural issues might include discussion of the bias in the referral process. A critical event in this stage would be the obtaining of parental consent for child evaluation.

The panelists in this group agreed that the technical assistance document would have to be modified by each state so as to be consistant with state requirements and monitoring procedures. It was suggested that a conceptual framework be available which the states could then develop into a technical assistance document. Even if the technical assistance document was developed by each state, the group felt that it was critical that the format allow school district personnel to add their own identified procedural issues, critical events, evaluation strategies, and description of successful and unsuccessful corrective actions. It was stated that unless the school districts could make it their own document, the technical assistance document would be little used.

Finally, the group did not recommend formal field-testing, although feedback to the developers of the framework was thought to be potentially useful. State education agency personnel and regional resource center personnel were named as possible disseminators of technical assistance documents.

Group II

Panelists in Group II recommended the development of a self-study guide which would address four major provisions of the law—protection in evaluation procedures, least restrictive environments, due process, and individualized education programs. Although they were concerned with the magnitude of such a guide, these panelists agreed that the components should not be viewed as isolated entities. In addressing themselves to the PEP section of the guide,

panelists suggested that the 4 position papers be used as a starting point in the development of the section.

The group also had specific suggestions on the content of the PEP section. First, recommendation was made that the PEP section not focus on an "average" school district, but try to address the evaluation needs of school districts at different levels of implementation. Thus, the section was conceived as addressing current compliance requirements, strategies for monitoring implementation for those districts in compliance with the law, and evaluation strategies. The evaluation strategies were described as directed towards quality issues and standards, with an emphasis on evaluating effects of implementation at the child level. The group also recommended that the evaluation component of the PEP section include the following subsections: (1) a rationale, (2) discussion of issues, (3) standards, (4) good practice examples, and (5) criteria, stated in measurable terms, for determining if standards have been met. The examples were viewed as providing illustration of alternative ways problems have been met, rather than indicating a one best practice.

Group II recognized the need for local school district adaptation of the guide, with the exception of compliance requirements. While these panelists did not recommend formal field-testing prior to dissemination, they suggested that a group of local school district personnel could be assembled for a "face validity check" of the guide. Feedback, subsequent to dissemination, was viewed as essential. The group recommended that the guide be revised based on the feedback of users such as school psychologists, parents, minority group representatives, and general administrators. Finally, the panelists were in agreement that development of the guide would require the expertise of specialists in packaging and dissemination.

Group III

Group III developed a model for implementing non-discriminatory procedures, rather than an evaluation guide. The model interrelates assessment and programming activities, and utilizes a cascade system with placements on a continuum ranging from less to more restrictive alternatives. As panelists explained the model, all children would begin in the least restrictive environment on the continuum, which is the regular classroom with no support services, but would continue filtering downward through successive placements on the continuum as needed. After regular classroom with no support services, successive alternatives on the continuum would be: consultant services to the regular classroom teacher, part-time tutorial help for the child, part-time resource room placement, full-time resource room placement, diagnostic/prescriptive class instruction (3-6 children for up to 6 weeks), self-contained classroom placement, and more intensive day or residential institutional placement.

According to the group, documentation would be required before a child would be transferred from a less to more restrictive placement alternative. Documentation would include description of the educational strategies which had been tried, as well as indication of the extent to which the strategies had been successful. No labeling of the child as a special education student would occur until or unless the child reached the fifth level of placement, full-time resource room. It is at this point that the child would be formally assessed by a school psychologist. Norm-referenced tests would be administered to document the fact that the child was handicapped. If this fifth level of placement was not successful, the child would be placed in a systematic instruction or diagnostic/prescriptive class. At this stage, assessment activity is tied closely to programming. The teacher would "systematically try to attain different kinds of objectives with the child, using different teaching strategies, different types of materials, and feedback." The goal is systematic evaluation of the child in order to determine how the child can best be taught.

Making the cascade system operational was viewed by the group as having several advantages over a more traditional placement and programming system. First, through-out the filtering process, adaptive behavior information is being collected. An individual child history is built which documents successful and unsuccessful education approaches in terms of child performance. Secondly, up to a point, a child can simply be viewed as an instructional casualty. Services can be delivered such as speech or remedial reading without additional labeling of the child. The system thus has the potential to be tied with compensatory education programs. Group III concluded that this approach, while not new to the special education field, would provide a base for fair decision-making, if implemented by school districts.

SUMMARY AND CONCLUSIONS

Commonalities among the subgroups can be summarized as follows:

1. Two of the three subgroups recommended that BEH disseminate its compliance procedure to local school districts. While recognizing that the Bureau's monitoring efforts are directed towards the state level of implementation, panelists thought the criteria would be of assistance to school districts in implementing the PEP provisions.
2. In general, panelists were quite concerned that another assessment technical assistance document would merely join administrators' shelves of other unused manuals on test selection and administration procedures. Pointing out that such "good practice" guides as the American Psychological Association's test standards have long been available, panelists felt that if technical assistance materials were to be used by school district personnel, they would require a format which encouraged user modification and individualization of the materials.

3. All groups emphasized the importance of tying assessment activities closely to information needed for referral, eligibility, placement, and programming decisions. The groups agreed that considering PEP or non-discriminatory testing as a discrete component of a sequentially ordered process would allow bias throughout the decision-making process.
4. All groups recommended that technical assistance documents or models be directed towards a local school district audience. While the groups varied in their emphasis on evaluation practices or implementation practices, two of the groups stressed the utility of identifying good-practice strategies that have been used successfully by school districts.

A major concern of the Bureau of Education for the Handicapped is the identification of children eligible to receive special education and related services. The position papers and summaries of panel discussions in this monograph suggest many strategies for improving both assessment procedures and the use of assessment information at the district and child levels of implementation. One strategy for improving assessment procedures on which most panelists agreed was the establishing of routine screening procedures at the school-district level. Routine screening procedures are encouraged by the Bureau as one means of insuring that children who may require special education services are referred for an individual evaluation. It is our hope that dissemination of this monograph stimulates other thoughts on achieving quality implementation of the Protection in Evaluation Procedures provisions of P.L. 94-142.

PROTECTION IN EVALUATION PROCEDURES CRITERIA STUDY PANEL PARTICIPANTS

May 10-11, 1978

Dr. Gordon Alley
Professor of Special Education
Lecturer in Pediatrics
University of Kansas

Ms. Jane Bauer
Consultant
Belmont, California

Dr. Louis Danielson
State Program Studies Branch
Bureau of Education for the Handicapped

Ms. Ann Geraghty
Division of Assistance to States
Bureau of Education for the Handicapped

Dr. Reginald Jones
Chairman and Professor of Afro-American Studies
Professor of Education
University of California, Berkeley

Dr. Mary Kennedy
Acting Branch Chief
State Program Studies Branch
Bureau of Education for the Handicapped

Dr. Richard Kicklighter
Coordinator, School Psychologists Services
Georgia State Department of Education

Dr. Jane Mercer
Professor of Sociology
University of California, Riverside

Dr. C. Edward Meyers
Consultant
Los Angeles, California

Dr. Linda Morra
State Program Studies Branch
Bureau of Education for the Handicapped

Dr. John B. Moyer
Assistant Professor of Special Education
Russel Sage College

Dr. Charles J. Murray
Special Education Department
New Jersey State Department of Education

Dr. Robert A. Oyler
Director of Special Education
Santa Ana Unified School District
Santa Ana, California

Dr. Ellis B. Page
Professor of Educational Psychology
University of Connecticut

Dr. Tom Salopek
Director of Special Education
Northwest Tri-County Intermediate Unit
Edinboro, Pa.

Dr. James Ysseldyke
Institute for Research on Learning Disabilities
University of Minnesota