DOCUMENT RESUME

ED 080 541 TM 003 046

TITLE	Invitational Conference on Testing Problems (New York, October 28, 1967).
INSTITUTION	Educational Testing Service, Princeton, N.J.
PUB DATE	28 Oct 67
NOTE	139p.
EDRS PRICE	MF-$0.65 HC-$6.58
DESCRIPTORS	Computer Assisted Instruction; *Conference Reports; *Curriculum Development; Individualized Curriculum; *Instructional Innovation; Intelligence Tests; *Measurement; Professional Education; Program Evaluation; Public Policy; Simulation; Statistical Analysis; *Testing Problems; Test Interpretation

ABSTRACT

The 1967 Invitational Conference on Testing Problems dealt with various aspects of change in education. Papers presented in Session I, Evaluation and Research in Curriculum Development, were: (1) "Adapting the Elementary School Curriculum to Individual Performance" by Robert Glaser, and (2) "An Evaluation Model for Professional Education--Medical Education" by Christine H. McGuire. Papers given in Session II, New Approaches to Instruction, were: (1) "Computer-Based Instruction in Initial Reading" by Richard C. Atkinson, and (2) "Academic Games and Learning" by James S. Coleman. The luncheon address was "Testing and Public Policy" by William Gorhan. Papers presented at Session III, Measurement Systems, were: (1) "Sample-free Test Calibration and Person Measurement" by Benjamin D. Wright, (2) "Reformation through Measurement in Secondary Education" by Paul R. Lohnes, and (3) "Surveys Undertaken by the Scottish Council for Research in Education" by David A. Walker. (KM)

PROCEEDINGS

of the

1967

Invitational

Conference

on

Testing

Problems

Library of Congress Catalog Number: 47-11220
Printed in the United States of America

Invitational Conference on Testing Problems

October 28, 1967
Hotel Roosevelt
New York City

BENJAMIN S. BLOOM
Chairman

EDUCATIONAL TESTING SERVICE
Princeton, New Jersey
Berkeley, California
Evanston, Illinois

Foreword

The 1967 Invitational Conference on Testing Problems focused upon various aspects of change in education. Speakers in the morning sessions were concerned with radically new approaches to instruction as well as the problems of evaluating changes now under way in curriculum programs in elementary and medical schools. In the afternoon session, we heard two speakers propose systems of measurement to bring about important changes in test calibration and grading and reporting procedures in secondary schools. The final speaker of the day, in a review of the surveys conducted by the Scottish Council for Research in Education, drew a profile of the changes that have taken place in Scottish education over the past three decades. All in all, it was an exciting program that reflected the progress and the problems that are part of the dynamic nature of education today.

We are all indebted to Professor Benjamin Bloom who, as chairman, was responsible for organizing this program. I should like also to extend our thanks to the luncheon speaker, Mr. William Gorham, and to the other speakers whose papers made this conference such a success.

Henry Chauncey
PRESIDENT

Among those who regularly attend the Invitational Conference is Professor E. F. Lindquist, co-founder of the American College Testing Program. He is shown above talking with Henry Chauncey (left), President of ETS. *Since the original group of 19 educators met in 1936, attendance at the conference has grown steadily. In 1967, 841 people attended. Teachers, psychologists, and others who participate in the Invitational Conference each year represent a wide range of interests and backgrounds.*

Preface

Having attended as well as participated in the Invitational Conference on Testing Problems for over two decades, I found it enlightening to see the conference from the viewpoint of the chairman. I began by asking for suggestions as to topics and speakers from participants in previous years. More than a hundred suggestions were made, and a number of persons volunteered their services as participants.

Since my own interests are in the relation between learning and testing, I attempted to select several speakers likely to focus on these relations. I then looked across the field to find persons whose work represented such advances in testing or education that they should be given an opportunity to be heard by the conference audience.

The papers by Glaser and McGuire deal with major programs of education at two very different levels of the educational system—elementary school and professional school. Both papers are concerned with the special role of evaluation in very new conceptions of instruction and learning.

The papers by Atkinson and Coleman deal with very different approaches to education—so different that testing is so much a part of instruction that separable evaluation procedures make little sense other than for research or demonstration purposes.

Quite in contrast is the paper by Lohnes, which suggests how a measurement system developed out of the vast store of data in Project TALENT could lead to very new types of education as well as guidance programs.

The paper by Walker is very useful in describing how the many surveys conducted by the Scottish Council for Research in Education evolved, and the impact they have had on the views of education and testing in that country.

In his paper, Wright demonstrates a radically new way of calibrating test items and test scores, which could have major consequences for the entire process of test construction and test interpretation.

Finally, the presentation at luncheon by Gorham raises the most fundamental issues of the relation between public policy and the field of educational testing.

It should be pointed out that the Invitational Conference is held in

such high esteem that each person invited to participate accepted as though a high honor had been conferred on him (or her). These are busy persons in great demand, and their responses to the invitation make it clear that the Invitational Conference provides an audience and a readership which stimulates the participants to their best efforts.

Once one has made the hard decisions about which speakers to invite, the support of the conference by Educational Testing Service makes the job of the chairman an easy one. Thanks to Anna Dragositz, this work goes so smoothly that the conference seems to be managing itself. Thanks are due to Henry Chauncey and ETS for the freedom given to the chairmen and for the support given to the conference over these many years. It is my fervent hope that the Invitational Conference may long continue to make its important contributions to education and educational testing.

Benjamin S. Bloom
CHAIRMAN

Contents

Session I

Theme:
Evaluation and Research
in Curriculum Development

Adapting the Elementary School Curriculum to Individual Performance*

ROBERT GLASER
University of Pittsburgh

Forty-two years ago, the twenty-fourth Yearbook of the National Society for the Study of Education was titled *Adapting the Schools to Individual Differences*. The first two paragraphs of Carleton Washburne's introduction (5) read as follows:

> The widespread use of intelligence tests and achievement tests during the past few years has made every educator realize forcefully that children vary greatly as individuals and that any one school grade contains children of an astonishingly wide variety of capacity and achievement.
>
> It has become palpably absurd to expect to achieve uniform results from uniform assignments made to a class of widely differing individuals. Throughout the educational world there has therefore awakened a desire to find some way of adapting schools to the differing individuals who attend them. This desire has resulted in a variety of experiments.

Four months ago, in the June 1967 issue of the *Review of Educational Research*, Nate Gage (4) pointed out that the contemporary arguments in favor of individualizing instruction are extremely plausible:

> . . . Learners do differ in ways relevant to their ability to profit from different kinds of instruction, content, incentives, and the like. Almost by definition, instruction adapted to these individual differences should be more effective.

*The research reported herein was performed pursuant to a contract with the Office of Education, U.S. Department of Health, Education, and Welfare. Contractors undertaking such projects under government sponsorship are encouraged to express freely their professional judgment in the conduct of the project. Points of view or opinions stated do not, therefore, necessarily represent official Office of Education position or policy.

> If so, why has not the evidence from attempts to individualize instruction yielded more dramatic results? Why are not the mean scores on achievement measures of pupils taught with due respect to their individual needs and abilities substantially higher, in unmistakable ways, than those of students taught in the conventional classroom, where everyone reads the same book, listens to the same lecture, participates in the same classroom discussion, moves at the same pace, and works at the same problems? For the fact is that, despite several decades of concern with individualization, few if any striking results have been reported.

One answer to the dilemma posed by Gage is the following: For the past 40 years, with few exceptions, studies of the outcomes of gross differences in educational method, while useful for immediate practical decision making in the schools, have been only superficially related to the kind of research required for obtaining cumulative reliable knowledge of the learning and teaching process. While studies often have been conducted with rigid experimental design, they have been conducted primarily to compare two or more different procedures. They have not been carried out in a way useful for building an organized body of information about the variables that influence learning by students in the classroom. A reasonable scientific methodology would have insisted that certain questions be posed about the procedures by which adapting to individual differences takes place in the classroom. One such question is this: How does a student, who is able to do so, take from classroom instruction what is suited to his needs; and more specifically, what relationships exist between individual capabilities and instructional methods that facilitate learning, achievement, and other expressed goals of good education?

It is obvious that we do not need studies which, at their outset, compare one educational procedure with another; what is required is the sustained design of educational environments in which the functional relationships between individual differences and learning method can be examined, and for which appropriate evaluative methodology can be developed. A primary prerequisite for such work is the establishment of school situations in which individual differences can be relatively freely adapted to, and from which extensive data can be obtained for, first, the analysis of individual difference measurements that are relevant to teaching practices, and, second, the assessment of the effects of alternate procedures for adapting to individual differences.

One of the programs of the Learning Research and Development Center at the University of Pittsburgh (an R & D center primarily

sponsored by the U.S. Office of Education) has been attempting to design procedures, materials, and an environment in an elementary school so that both research and development can take place on the process of adapting to individual differences. Over the past few years, under the direction of John Bolvin, the initial procedure, which is under constant revision, has come to be called "individually prescribed instruction." In this paper I should like to describe our beginning approach.

Requirements for Individual Progress

In general, it is assumed that certain requirements for adapting to individual differences have to be met for the design of an individualized system. These requirements are the following:

1. The conventional boundaries of grade levels and arbitrary time units for subject-matter coverage need to be redesigned to permit each student to work at his actual level of accomplishment in a subject-matter area, and to permit him to move ahead as soon as he masters the prerequisites for the next level of advancement.

2. Well-defined sequences of progressive, behaviorally defined objectives in various subject areas need to be established as guidelines for setting up a student's program of study. The student's achievement is defined by his position along this progression of advancement.

3. A student's progress through a curriculum sequence must be monitored by adequate methods and instruments for assessing his abilities and accomplishments so that a teaching program can be adapted to his requirements.

4. Students must be taught and provided with appropriate instructional materials so that they acquire increasing competence in self-directed learning. To accomplish this, the teacher must provide the student with standards of performance so that he can evaluate his own attainment, and teaching activities must be directed by individual learner accomplishment.

5. Special professional training must be provided to school personnel so that they can accomplish the evaluation, diagnosis, and guidance of student performance that is required to organize instruction for individualized learning—as contrasted to the total-class management of learning.

6. The individualization of instruction requires that the teacher attend to and utilize detailed information about each student in order to design appropriate instructional programs. To assist the teacher in processing this information, it seems likely that schools will take advantage of efficient data processing systems.

The technicalities for designing and implementing a system with these requirements and the necessary teacher, administrative, and materiel needs are demanding. The questions involved in measuring individual differences in learning and performance, making adequate student diagnoses, building appropriate learning materials, and matching student differences to instructional alternatives need to be formulated and answered. With this in mind, I shall discuss some major aspects and indicate some of the questions that have been raised in our attempts to individualize an elementary school curriculum.

Definition of Educational Objectives

First, some things need to be said about the analysis and definition of a continuum of educational objectives. While the objectives of one curriculum designer may not be another's, one of the most important factors that can contribute to improvement in educational attainment in an individualized system is the analysis and specification of the desired outcomes of learning. In the interest of brevity, the following points concerning this first step are made without elaboration:

1. The definition of instructional objectives instructs the curriculum designer and the teacher how to proceed. Vague specification of the desired competence level leaves the teacher with little concrete information about what to look for in student performance and about what to provide to the student to attain or surpass this performance.

2. The interaction between the specification of objectives and experience in teaching frequently provides a basis for a redefinition of objectives. The process of clarifying goals, working toward them, appraising progress, reexamining the objectives, modifying the instructional procedures to achieve goals, and clarifying the objectives themselves in the light of experience and data should be a continuous process.

3. Regardless of the way a subject matter is structured, there is usually present some hierarchy of subobjectives indicating that certain per-

formances must be present as a basis for learning subsequent tasks. Absence of the specification of prerequisite competence in a sequence of instruction dooms many students to failure.

4. A student's knowledge of objectives gives him a goal to attain; such knowledge is instructive and motivating. It permits the student to monitor his partial successes and failures and to adjust and organize learning resources for himself.
5. As in other lines of endeavor, teachers require frequent information about the results of their work so that they can adjust their practices accordingly. Teachers need standards by which to judge themselves and by which society can judge their effectiveness.
6. The exercise of specifying objectives points up the inadequacies and omissions in a curriculum. The fear of many educators that the detailed specification of objectives limits them to only simple behaviors which can be forced into measurable and observable terms is an incorrect notion. If, indeed, complex reasoning and open-endedness are desirable aspects of human behavior, then this needs to be a recognized and measurable goal. Overly general objectives may force us to settle for what can be easily expressed and measured.

In our project (also frequently referred to as the "Oakleaf Project" after the name of the elementary school in the Baldwin-Whitehall school district in suburban Pittsburgh where we started the procedure of individually prescribed instruction), the kindergarten through sixth-grade mathematics curriculum has identified 430 specific instructional objectives. These objectives are grouped into 88 units. Each unit comprises an instructional entity that the student works through at any one time; on the average, there are 5 objectives per unit, with a range of 1 to 14. A set of units consisting of different subject areas in mathematics comprises a level; levels are labeled with letters A through H, and can be thought of as roughly comparable to a school grade level. Table 1 provides a content outline of the organization of the curriculum units. In a revised version of this curriculum to be studied during the 1967-68 school year, we have been assisted in the preparation of tests and instructional materials by Appleton-Century-Crofts.

Assessment and Diagnosis

A second major requirement in an individualized program is assessment and diagnosis of student performance so that the amount and kind

Table 1

Description of Selected Mathematics Curriculum Units

Unit No.	Level	Unit Label	Short Description	Approximate Conventional Grade Level
1	A	Numeration	Counting to ten.	
2	A	Addition	Addition to sums of six.	
3	A	Fractions	Identification of 1/2 of sets.	1
4	B	Numeration	Counting to 100; ordinals to 10th.	
5	B	Addition	Addition to sums of 12.	
⋮				
11	C	Numeration	Counting and skip counting to 200.	
12	C	Place Value	Recognizes place values and concepts of "greater than; less than."	
13	C	Addition	Two-digit sums without carrying.	2
14	C	Subtraction	Two digit differences without borrowing.	
15	C	COP*	Selection of operation to solve problems.	
⋮				
23	D	Numeration	Counting and skip counting to 1,000.	
24	D	Place Value	Makes place value charts to thousands.	
25	D	Addition	Begins addition with carrying.	
26	D	Subtraction	Begins subtraction with borrowing.	
27	D	Mult'	Begins multiplication as repeated addition with factors to 5.	3-4
28	D	Division	Begins division as partition with divisors to 5.	
29	D	COP*	Problems requiring many processes.	
⋮				
37	E	Numeration	Identifies odd and even numbers; converts mixed decimal fractions.	
38	E	Place Value	Place value to billions; begins exponents.	
39	E	Addition	Addition with carrying to 4 digits.	
40	E	Subtraction	Subtraction with borrowing to 3 digits.	4-5
41	E	Mult'	Uses associative and distributive principles and does simple multiplication with carrying.	
42	E	Division	Uses ladder algorithm for division	

*COP stands for Combination of Processes
'Mult. stands for Multiplication

Table 1

(Continued)

Unit No.	Level	Unit Label	Short Description	Approximate Conventional Grade Level
43	E	COP*	Solves problems using n as variable.	
⋮				
51	F	Numeration	Rounds numbers; identifies prime numbers.	
52	F	Place Value	Manipulates exponents to ten cubed.	
53	F	Addition	Adds large sums to seven digits.	
54	F	Subtraction	Subtracts to seven digits.	5-6
55	F	Mult'	Multiplication with 3 digits.	
56	F	Division	Division algorithm with no remainders; simple division with remainders.	
57	F	COP*	Performs multiple operations with number pairs.	
⋮				
65	G	Numeration	Uses prime numbers to factor composite numbers; operations in bases 5 and 10.	
66	G	Place Value	Charts numbers by place value in base 5.	6---
67	G	Addition	Adds positive and negative numbers.	
68	G	Subtraction	Subtracts negative and positive numbers.	
69	G	Mult'	Multiplies numbers in exponential form.	
⋮				
76	H	Numeration	Identifies numerals in base 2, 3, and 8.	
77	H	Place Value	Place value charts in other bases.	
78	H	Add & Sub--Other Bases	Adds and subtracts in bases 2, 3, 5, 8.	
79	H	Addition	Adds with negative powers of ten.	
80	H	Subtraction	Subtracts with negative powers of ten.	7---
81	H	Mult & Div--Other Bases	Multiplies and divides in bases 2, 3, 5, 8.	
82	H	Mult'	Multiplies with decimals and negative numbers.	
83	H	Division	Divides decimal numbers, positive and negative numbers; calculates square roots.	
84	H	COP*	Solves word problems with skills learned.	
⋮				
88				

of instruction can be adapted to the student's particular requirements. From this point of view, testing and teaching are inseparable aspects and not two different enterprises, as one might be led to believe by current practices in education. Frequent information about student performance is used as the basis on which the teacher decides on the next instructional step; and equally important, it also serves as feedback to the student. It is also invaluable for the design and redesign of teaching materials.

The kind of measurement required for these purposes forces a distinction between performance measurement and aptitude measurement. The instruments used to measure performance are specifically concerned with the properties of present behavior as they relate to the requirements for deciding on subsequent instructional steps. It seems easier, in a sense, to predict the next moment in time in a lesson sequence than to predict long-range performance, which is a task usually set for aptitude measurements. It is possible that measures predictive of immediate learning success are different from those employed for more long-range prediction. Some of the factor studies of changing ability constellations over learning (3) suggest that this may be the case.

The testing procedure so far designed under the direction of Richard C. Cox, with the evaluation support of C. M. Lindvall, is oriented toward subject-matter mastery. For every unit in mathematics there is a pretest and a posttest. A pretest samples the various objectives in the unit and is diagnostic enough to pinpoint mastery or the lack of it in the various component skills. A posttest assesses the material that a student has been taught and is essentially an alternate form of the pretest. For each objective within the unit there is a curriculum-embedded test that is part of the instructional sequence. These curriculum-embedded tests not only measure performance on the objective on which the student has been working but also include test exercises on the next objective that the student is likely to work on. The notion here is that if a lesson is taught well, the student will learn not only the present lesson, but will be able to master exercises in the immediate subsequent skill. It is a special challenge for lesson writers to make this "testing out" of an objective as frequent an occurrence as possible.

At the beginning of a school year, a student takes one or more wide-band placement tests which consist of sample items measuring his mastery of the objectives of each of the units within a level of work. On the basis of his last year's performance, an approximation is made of the student's level of achievement, and testing begins from there. The

student is tested over a range from what he knows to what he has not yet learned. Depending on his background, and depending to some extent on how hierarchical the achievement objectives in a subject matter are, the student's performance may be more or less cumulative. In the first year at Oakleaf, the achievement assessed by the placement tests was spotty—that is, students showed areas of mastery and areas of weakness at various places along the continuum The data from the placement tests in subsequent years show a more cumulative pattern of achievement, which may be a result of the individualized curriculum.

Tests are seen as part of instruction, and the students look forward to them because they get immediate information about whether they need additional work in a unit or can move on to new work. The overall philosophy of this built-in testing program is that at any point in time the student's performance is so monitored that a detailed assessment is available of his performance and progress. The continuous recording and updating of these performance data seems to make special testing procedures unnecessary. As we get better in designing a curriculum which adapts to individual differences, I suspect that the test-taking aspects generally present in education will diminish, as perhaps will the test-anxious or test-sensitive student.

Consider now some of the problems in testing that arise and require investigation: One point is that initial placement tests take an undue amount of time to administer, especially to new students entering an individualized program. Some form of sequential testing should be helpful. An interesting idea is the use of a computer terminal for such testing. A second point that is more fundamental, however, is the problem of analysis of the dimensions of individualization that can be measured and are useful for instructional decision making. At the present time, the measures obtained in the mathematics curriculum are measures of achievement in the various units and objectives, with some further indices of the rate at which a student has been achieving mastery and the amount of practice and review he has required. Little use is made, at present, of measures of general intelligence or aptitude which have seemed difficult to relate to instructional decisions in the elementary school. From the placement tests no measures are obtained of subtle aspects of learning style, but perhaps reliable measures of this can be found. It is our general contention that the most useful measures of learning characteristics related to instructional decisions will result from indices obtained from monitoring the student's learning characteristics and performance over a period of time in the curriculum.

As a student goes through the units in the mathematics curriculum, a posttest mastery criterion of 85 percent is employed—that is, a student must achieve this level of performance before he moves on to the next unit. The setting of a criterion level, however, is an experimental question which needs investigating. Assuming a reasonably cumulative curriculum where new learning depends upon previous learning, do different units and differing students require a uniform level of proficiency? If too high a criterion is set, a student can spend too much time mastering fine points of one unit, while he might be beginning the next. A bright student might begin to learn multiplication while still becoming proficient in the fundamentals of addition and subtraction, and in this way develop a richer concept of addition; another student may require more detailed mastery of fundamentals before he moves on. The questions involved seem more complex than we had originally supposed.

Data Management

The accumulation and maintenance of the day-to-day records required for individualized instruction is a sizable enterprise for a school. In the initial years of the Oakleaf Project, we have been accomplishing this by hand. Each teacher has the assistance of an aide for individualized classes, and there is a data processing room with a staff of clerks who receive information from teachers and teacher assistants, process it, and return it to the classroom. After using this simulated computer system for two years, we designed an initial computerized data processing system. At the present time, in cooperation with the General Learning Corporation, we are investigating a computer management system to assist in researching and implementing individualized instruction. In its initial operation, there is a teacher terminal at the school which the teacher can interrogate for information; there is a terminal back in the laboratory which can be used to write programs, to analyze various aspects of student performance, to try out various data-reduction routines, and to analyze the instructional effectiveness of various curriculum units so that they can be revised when necessary. The particularly challenging research aspect of a computer-management system is the task of matching relevant measures of student performance with appropriate curriculum methods and materials to provide the teacher with assistance in preparing instructional prescriptions for each student. More needs to be said on this, but first another aspect must be mentioned.

Learning and Teaching

A system for adapting to individual differences requires more than specification of objectives, measurement and assessment of these objectives, and the monitoring of student performance and progress. It also requires learning and teaching. The primary task to be faced here is that instruction involves teaching to the student and not to the classroom group. This has created a problem for many teachers who have been trained to teach a class and have had much less experience in teaching individuals. Another problem is that instructional materials, especially in the elementary school, consist of texts and workbooks designed to be used with group directions and group-paced exercises.

Adapting instructional materials and procedures to individual differences is a function of both student behavior and the nature of the subject matter being taught. It is important to emphasize at this point that individualization is accomplished by designing a particular curriculum for the needs of a student (the word "needs" is used operationally in terms of student characteristics that we can reliably assess and that are relevant to instructional decisions). Adapting to individual requirements does not at all imply that a student necessarily works alone or in any particular mode or setting. In the course of individualized instruction, students may be taught by lecture, by programmed texts, by group discussion, by group projects, or by teaching machines. The essential notion is that individual requirements are matched to appropriate instructional procedures. For much of mathematics, a self-instructional situation may be suitable; it may be less suitable for various components of the social studies and the language arts. The individualization of instructional procedures certainly involves a variety of modes of learning. (Perhaps the term "individualized progress" is less misleading in this regard than the term "individualized instruction.")

In the elementary school, general education curriculum objectives are more or less the same for all students so the differentiation of learning goals may not be an appropriate procedure for adapting to individual differences. While the goals are the same, however, the pattern of the specific subgoals may differ to the extent that different students may work through a sequence of different topics to reach the same goal. In this way, some adaptation can take place by individualizing instructional tasks. Individualization also takes place by allowing for different learning rates involving different amounts of repetition and materials which permit smaller or larger instructional steps. These two

modes of individualization are the easiest to implement on the basis of student performance. Other modes of adapting to individual differences involving different media and different instructional methods are more difficult to implement because we know very little about the relationship between measures of student behavior and the learning effectiveness of these various means and media of instruction.

A basic principle in designing instructional materials and environments for individualized learning is to provide situations that are highly responsive to the behavior of the student. It is well known that learning occurs because the learner acts on his instructional environment, changes it, and is changed in turn by the consequences of his actions. As learning proceeds, new consequences in the environment are established with which the learner interacts. It is the management of the contingencies between student performance and environmental change that is the fundamental task of the teacher and the tools with which he is provided. This intimate dynamic relationship has been the aspiration of the work in programmed instruction and should be the goal of systems for individualizing instruction. An environment highly responsive to the student's endeavors seems to be capable of resulting in the effective attainment of competence, but is also motivating in the sense that it reinforces the kind of behavior that is alluded to when we use phrases like "developing a sense of exploration and curiosity," "a sense of inquiry," and "a sense of control over one's own education." To date at the Oakleaf school, teachers have been provided with an initial set of materials and procedures from which they can select in prescribing a student's instruction. By selecting these instructional means on the basis of the student's performance record and general behavior, the teacher, in essence, can make up a unique set of activities for each student. The decision process intervening between student assessment and the assignment of instructional activities is the essential task of individualized instruction that needs to be studied and understood.

Certain things seem easy enough to do—at least relatively easy. For example, a student's entering behavior can be assessed in order to diagnose his subject-matter competencies. Some attempts can be made to determine the kinds of materials in which he would be interested as a result of his general background. It is more difficult to match differential aptitude patterns to learning procedures, if indeed the usual kinds of aptitude measures are at all relevant for this purpose. Most promising, as has been suggested, are the measures that can be obtained after the performance of the individual student has been monitored for some

period of instruction. Indices obtained from a detailed record of student performance should provide good information relevant to subsequent learning requirements. The fascinating question for research is the following: Given the properties of a particular subject matter, relevant measures of individual differences, available instructional means, and specification of desired learning outcomes, what are the relationships between these variables that provide for optimal learning conditions for an individual student? This question can be studied in a number of ways (2), but discussion of these methodologies requires another meeting.

In our work, the teacher, provided with information about the student's progress and about available resources, prepares an instructional prescription for the student. In the future, it is conceivable that the teacher might be presented with suggested alternative student prescriptions which she can accept, reject, or modify; certain prescriptions could be presented to the student directly. At the present stage of our knowledge, the decision rules for going from measures of student performance to instructional prescriptions may not be very complex, but little is known about the amount of complexity required, although the individual monitoring of student performance provides us with a good data base to study this process. Study of attempts at individualization should point out how fine or coarse adaptation to individual differences can be with the knowledge at our disposal. Will it be possible to make unique prescriptions for each individual, or will it be discovered that, in an elementary school with a certain population, instruction can be quite effective by having, at each decision point, three to ten instructional alternatives? This may provide all the variability required or that can be produced. In time, the order of complexity may be like that employed in medical diagnosis and will be as crude or as sophisticated, depending on one's point of view, as medical diagnostic and treatment relationships.

Changes in Classroom Communication Structure

In the work to date on individually prescribed instruction, we have designed a situation in which it seems possible to study the why and how of adapting to individual differences in achievement in the elementary school. First, we can examine changes in classroom communication structure which result when the change is made from teacher control of a total class to a procedure that attempts to individualize

instruction. A study to investigate this was part of a larger study carried out by Hilton Bialek of the George Washington University HumRRO Unit in California when individually prescribed instruction was started in four schools (1). The general hypothesis Bialek set forth was that once an individualized program was introduced, change would occur in the initiation of communication between teacher and student and in the relative distribution of instructional and noninstructional communications. Twenty-one experimental classes in four schools were observed, once before the introduction of the individualized program and four times after the program started; control classes were observed three times throughout the course of the school year. Six categories of teacher-student communication were used for recording observations: teacher to one student, teacher to more than one student, and student to teacher; each of these was subdivided into instructional or noninstructional communications. "Instructional" was defined as describing or explaining an instructionally relevant procedure, conveying instructional information, or raising a problem for discussion; "noninstructional" was defined with a non-negative connotation as social exchanges unrelated to the subject matter at hand, discipline or class control, or evaluation inquiries as to whether the student knew or retained information.

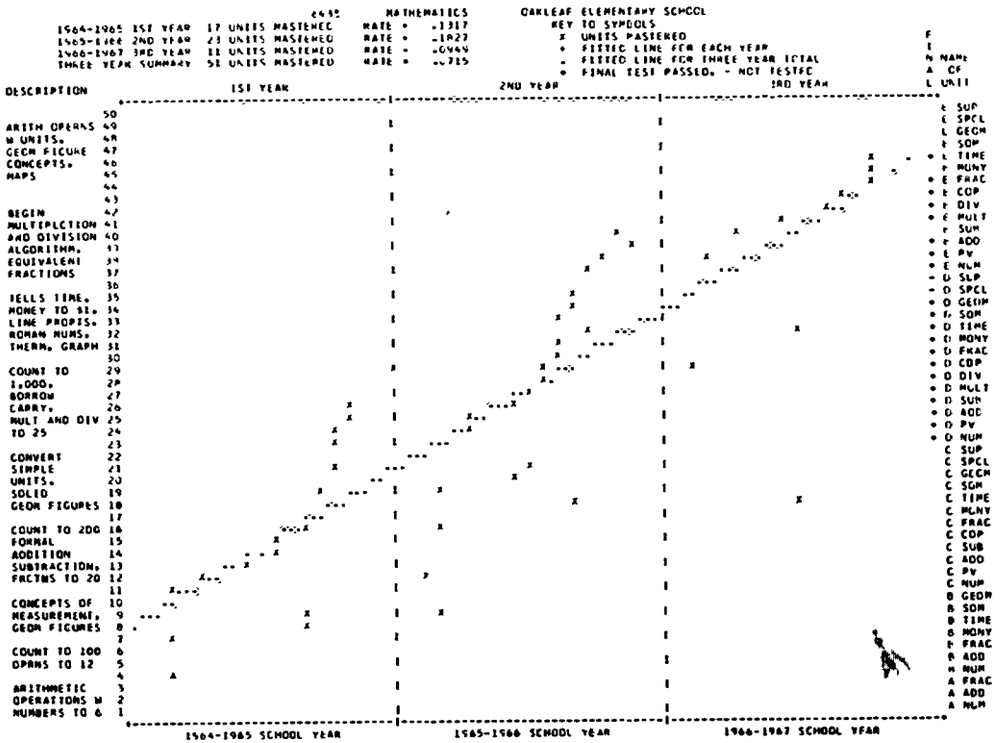
In the control classes, three aspects of the communication pattern appeared as follows: 1. Over half of the communications in the classroom were noninstructional; 2. about 90 percent of the communications were teacher-initiated; half of these were directed to the single student and half to groups of students; and 3. when the teacher talked to one student, it was most likely that the communication was noninstructional; when the teacher talked to more than one student, it was likely that the communication was instructional. Before the initiation of the individualized program, the communication pattern in the experimental classes was highly similar to this control-school pattern. After the introduction of the individually prescribed instruction procedure, the following appeared: 1. Over three quarters of the communications were instructional in nature; 2. 20 percent of the communications were teacher-initiated; of these, three quarters were directed to the single student; 3. about 80 percent of the communications were student-initiated; of these, three quarters were instructional in nature; and 4. there was a trend for the overall number of communications to decrease in the experimental classes. In general, in the individually prescribed instruction classroom, it was clear that the responsibility for

teacher-student communication fell upon the student, and that the content of most communications was instructional in nature. Bialek pointed out that student-to-student communication was not recorded but that much of it occurred during the experimental classes.

Patterns of Student Progress

What patterns of student progress occur under a system of individualized instruction? Figures 1 to 13 present computer-plotted summary charts that show the progress of different students over three years of school in the mathematics curriculum at the Oakleaf school. In Figure 1, the vertical axis on the left-hand side lists the numbers of the curriculum units. There are 88 units in the curriculum sequence; on each chart, these unit numbers may or may not begin at 1 or end at 88, depending on the level and unit at which the student originally placed. Very general descriptions of sets of units are given along this axis to show what the student is working on. For example, around unit 40, a student would be working on beginning multiplication and division algorithms and on equivalent fractions. The vertical axis on the right-hand side shows the same thing, but lists the levels A through E and the names of the units in the level. As I have indicated, roughly compared with standard textbooks, A is kindergarten, B is first-year work, C second-year work, D third and fourth year, E fourth and fifth, F fifth and sixth, and G sixth and above. On the horizontal axis, units that are mastered during a particular two-week period over the three years of school are plotted. Every time a unit is mastered, an X is plotted. The Xs represent a rather stringent mastery criterion of 85 percent, and are only plotted on this chart if such a mastery level has been attained either on a unit posttest or a pretest. An X is also plotted when a student requires review and repeats some work in a unit in order to reattain proficiency. Teachers use the 85 percent mastery criterion as a basis for prescribing new work; sometimes, however, they decide that a student should be permitted to go on without insisting that he meet this criterion, but this is not shown here. The number of units mastered is one measure of student rate through the curriculum, although units differ widely in the average time required to work through them. The average time to master a unit is 12 days with a range of 1 to 60 days, one day representing pretest mastery. In Figures 1 to 13, the patterns of Xs show how achievement progresses for different students. The straight line of dots on the chart represents a linear least squares fit of

Figure 1 *Janie*



the Xs to represent a general rate of progress over three years.

In Figure 1, Janie, a third grader, has, over her first three years of school, worked up to unit 46 and worked on 51 units to the 85 percent criterion (including the repetition of units reviewed). In Figure 2, Leonard, her classmate, worked up to unit 40 and worked to criterion on 40 units. Janie began in the first year at levels B and C, whereas Leonard spent much time in his first year reviewing work in levels A and early B, which is elementary arithmetic operations. These two relatively swift students are compared with their classmates Jimmy and Joey shown in Figures 3 and 4. These students worked to criterion on 19 and 20 units respectively over the three years and worked up to units 20 and 18. Look at Figure 1 again. The bullets on the right-hand vertical axis show the result of a test, prepared for the level or levels at which a student worked during the third year. Each bullet represents high mastery and retention—85 percent; a blank space indicates less than 85 percent; a dash means that the test was not given on that unit. Looking at the bullets, the two swift students shown in Figures 1 and 2 show excellent mastery, especially Janie. For the two slower students Joey showed somewhat more final mastery than Jimmy. In Figure 5, Phyllis has impressed her teachers by an increasing rate of achievement, having shown proficiency on 4, 10, and 14 units in each of the three years, 28 units in all. Phyllis started out in the first grade like Jimmy, but in the third grade she covered twice as many units as he did.

In Figures 6 and 7, we see Charles and John in the fourth grade, two fast top students, showing proficiency in 46 and 53 units respectively over the three years, but showing a somewhat different pattern. In fourth grade, John moved very fast, but required a significant amount of time in review work, as so decided by the teachers in their prescriptions. Charles showed a more cumulative pattern of achievement requiring less review and a more consistent pattern of retention. In Figure 8, a slow student, Ralph, covering 26 units, shows an idiosyncratic pattern of achievement where mastery of many units occurs at a certain time of the year. Ralph's teachers say that he shows fluctuating motivation and finds it difficult to work steadily for periods of time.

Figure 9 shows that Bruce, a fifth grader, covered 65 units over the three years (the computer printout had space for only 57 units). At the beginning of each year, he required some review of the previous year's work, which he accomplished rapidly. It is of concern to us that high proficiency requirements for this review may be slowing the student

Continued on page 28

Figure 2 *Leonard*

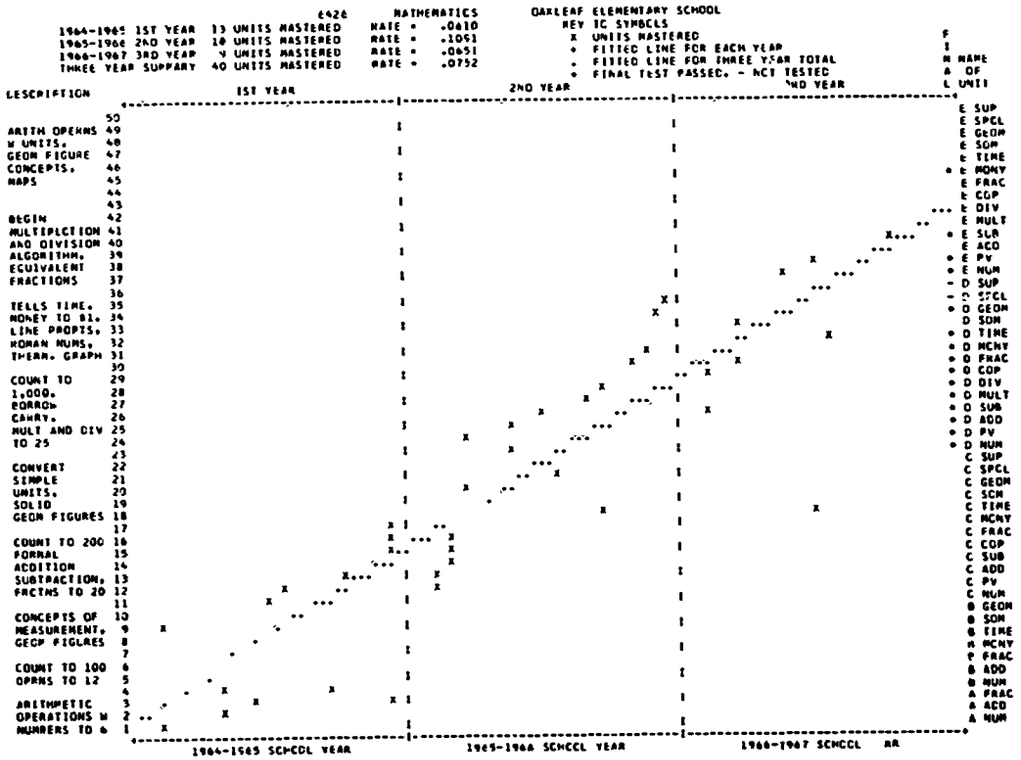


Figure 3 *Jimmy*

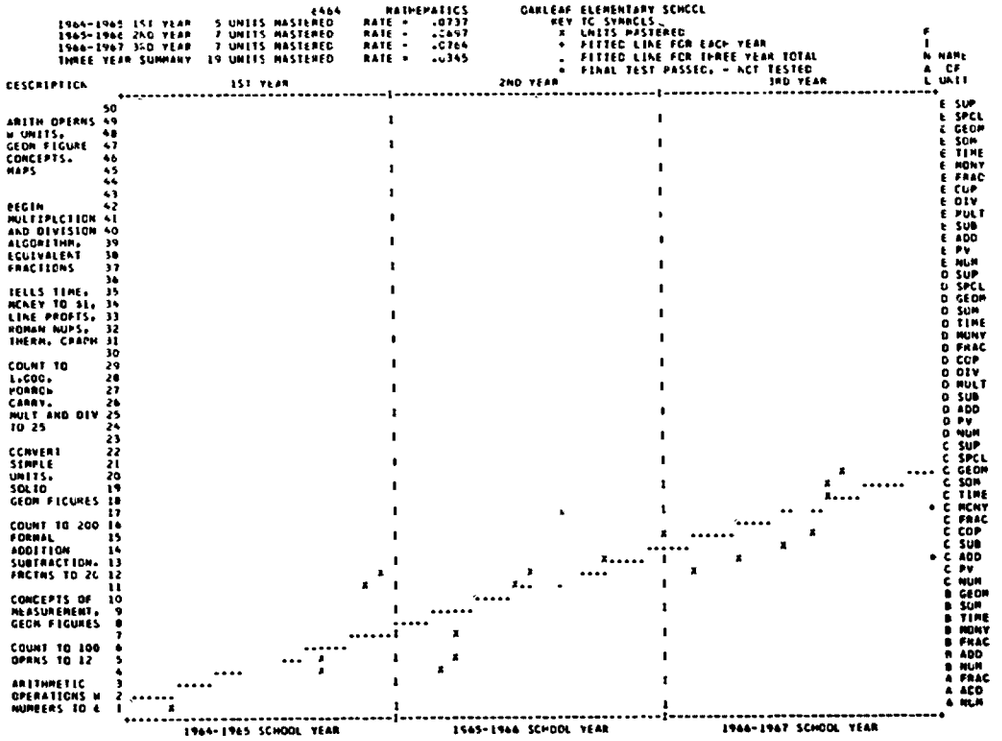


Figure 4 *Joey*

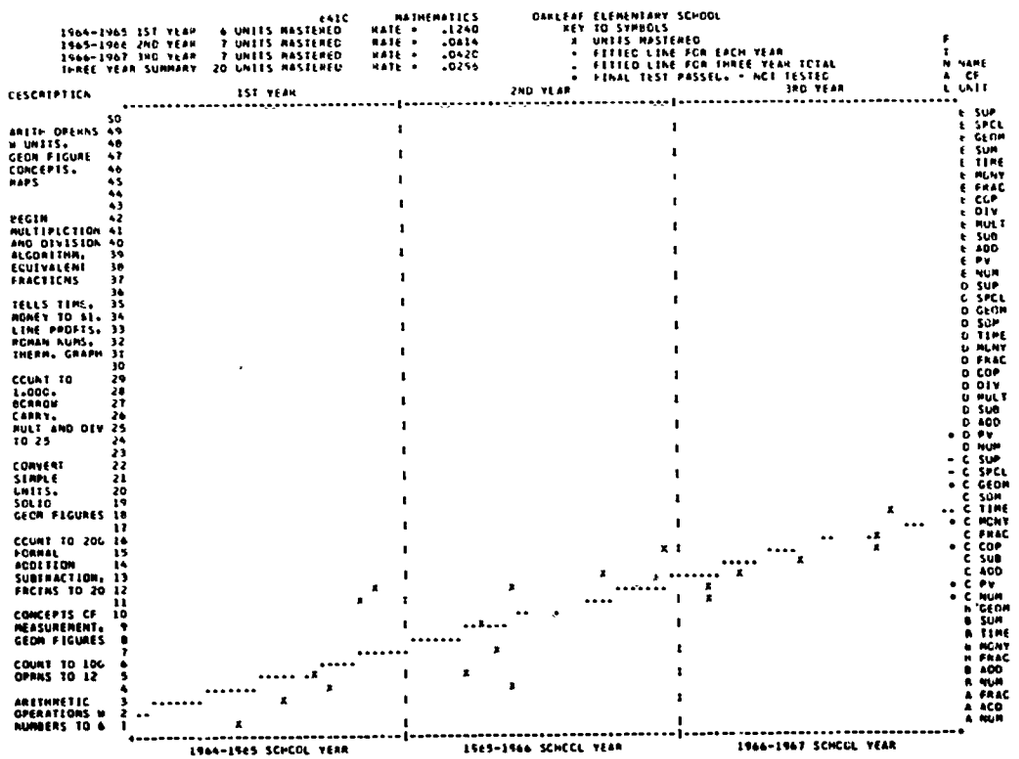


Figure 6 *Phyllis*

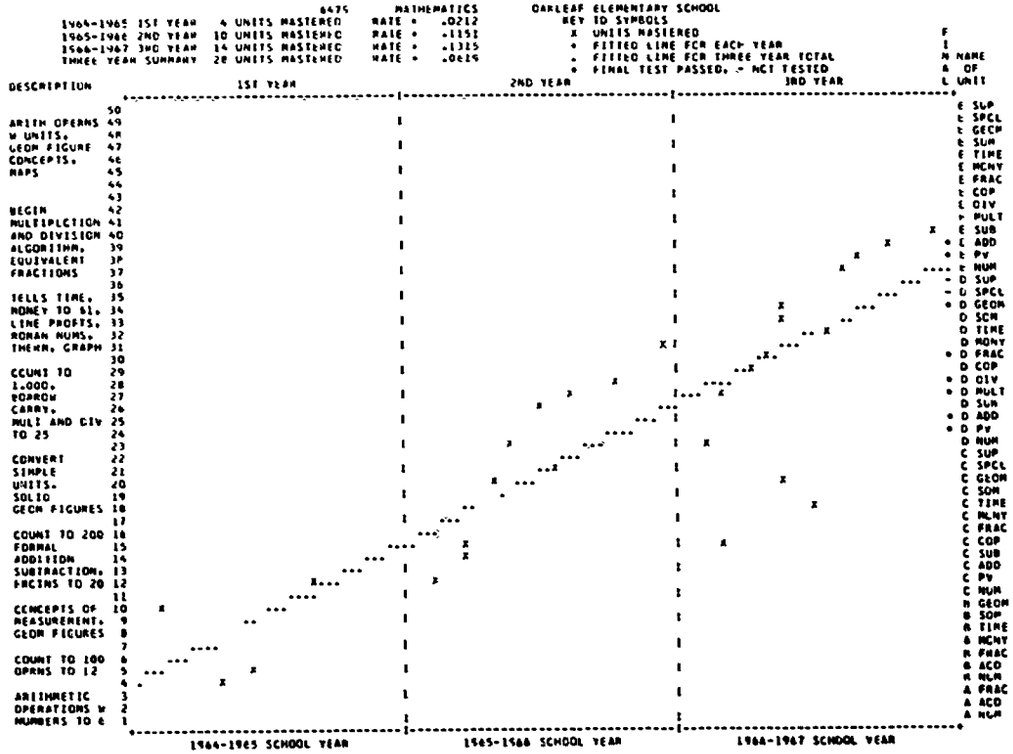


Figure 6 *Charles*

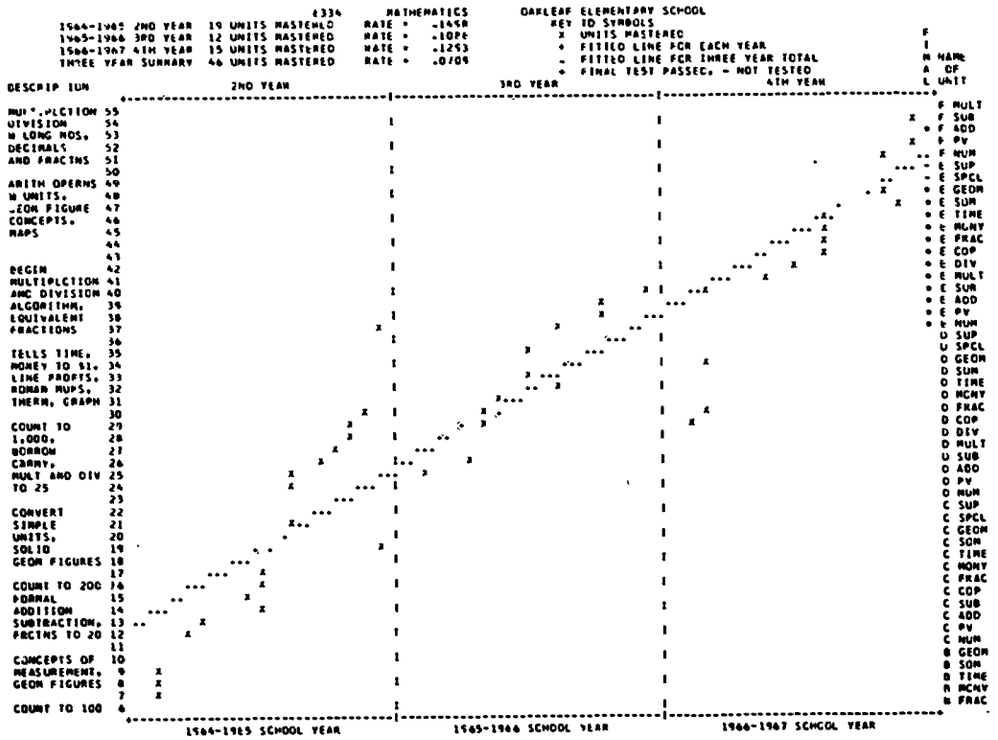


Figure 7 *John*

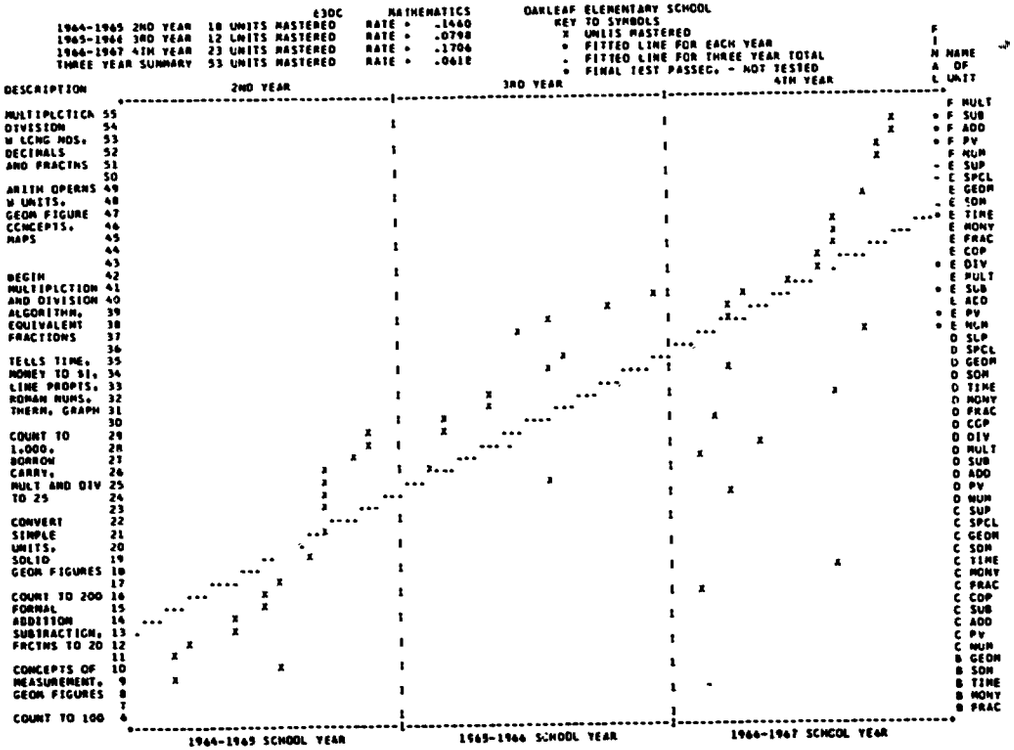


Figure 8 *Ralph*

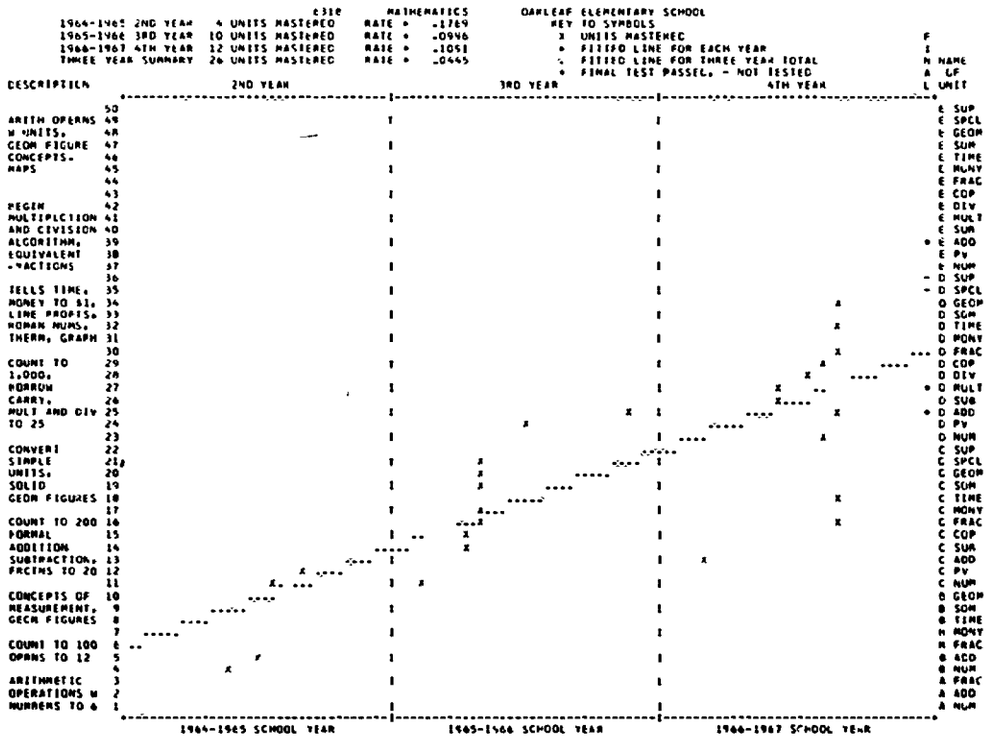
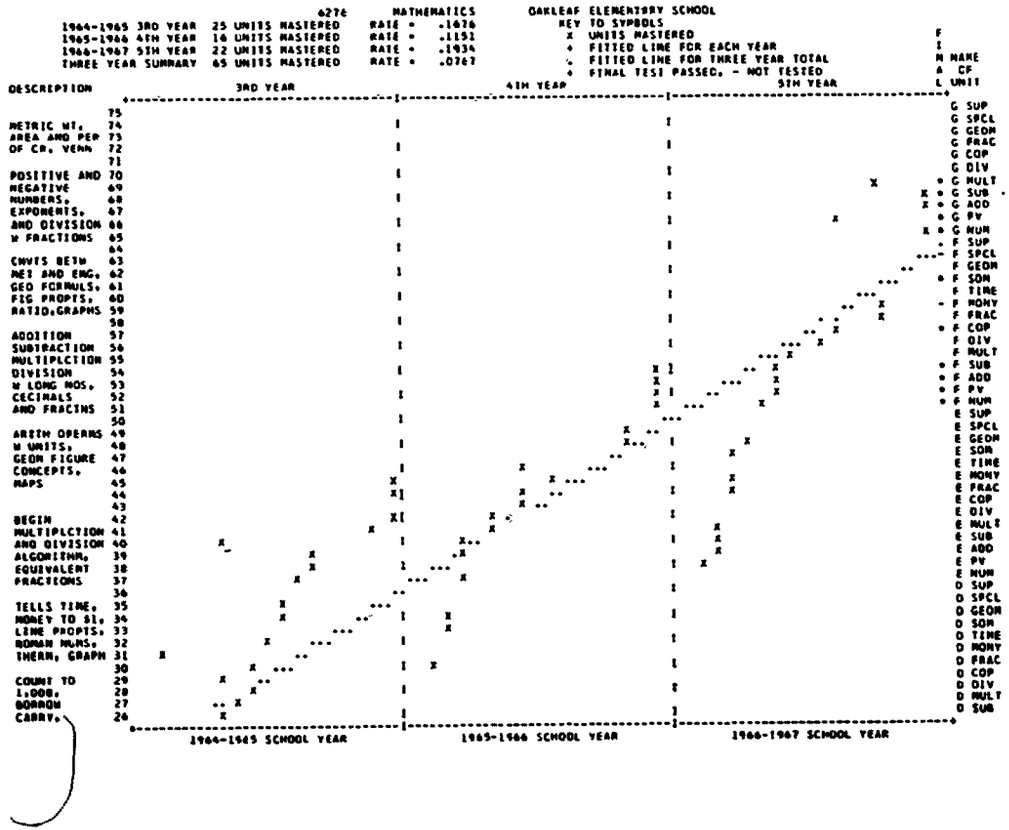


Figure 9 *Bruce*



down, and studies are under way to investigate the required characteristics of review and practice. Bruce is in the fifth year of school and is working on topics generally taught in the late sixth grade and beyond. In Figure 10, Allen shows half the rate of Bruce, Allen working to about the 30th unit and Bruce working to the 67th unit. Figure 11 shows Timothy finishing the fifth grade at the 42nd unit. His brother, in the third grade, ended up at the 39th unit.

In the sixth grade, Charlene (Figure 12) covered 66 units over the three years, working to unit 72; her progress shows much quick review, particularly in the first month of sixth grade. Lack of retention over the summer vacation is apparent. In Figure 13, Diane is seen as a slower sixth-year student. Diane showed a non-cumulative remedial pattern before she showed any cumulative advancement. Patterns similar to Diane's occur frequently in the shift from class to more individualized instruction.

For the 100 students who have been at Oakleaf for three years, the mean number of units mastered over these years is 37 with a standard deviation of 12; the maximum number of units covered by a student is 73 and the minimum is 13, a range of 60 units. It appears that the number of units covered increases in the higher years of work. This may be a function of either the nature of the units at the higher levels of the curriculum, the ability of the older students to move faster with our materials (implying that we might do a better job at the earlier levels), or an artifact of the amount of review assigned by the teachers at the beginning of the year, which would increase the number of units mastered.

One indication of the consistency of the rate with which a student moves through the mathematics curriculum is given by the correlation of the number of units covered in the different years of the program. The three intercorrelations between the first, second, and third years are .37 (1 vs. 2), .45 (2 vs. 3), and .52 (1 vs. 3). These are not as high as one would expect. The correlation between the knowledge that a student brings with him at the beginning of his first year—that is, where he places and begins in the curriculum—and the number of units he covers over three years is .61; this relationship is also reflected by a correlation of .72 between the number of the unit begun in the first year with the number of the unit reached at the end of the third year. The correlation between the rough IQ measure used by the school system, the California Test of Mental Maturity, and total number of units covered over three years is .32; similarly, the correlation between IQ and the number of the unit reached at the end of the third year is .31.

Figure 10 *Allen*

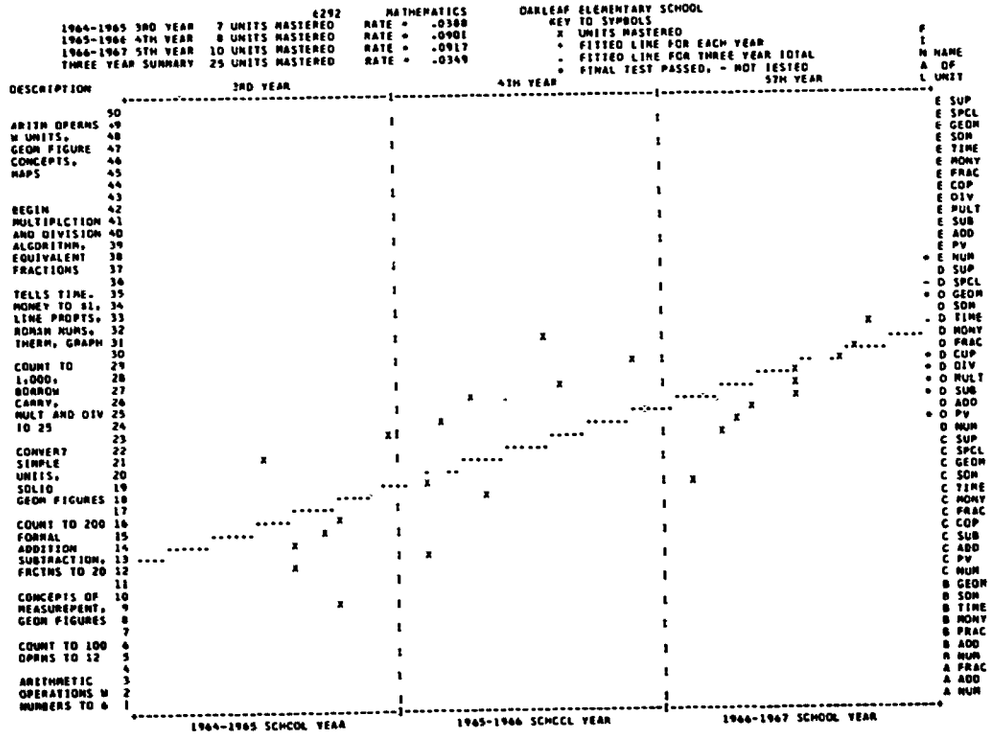


Figure 11 *Timothy*

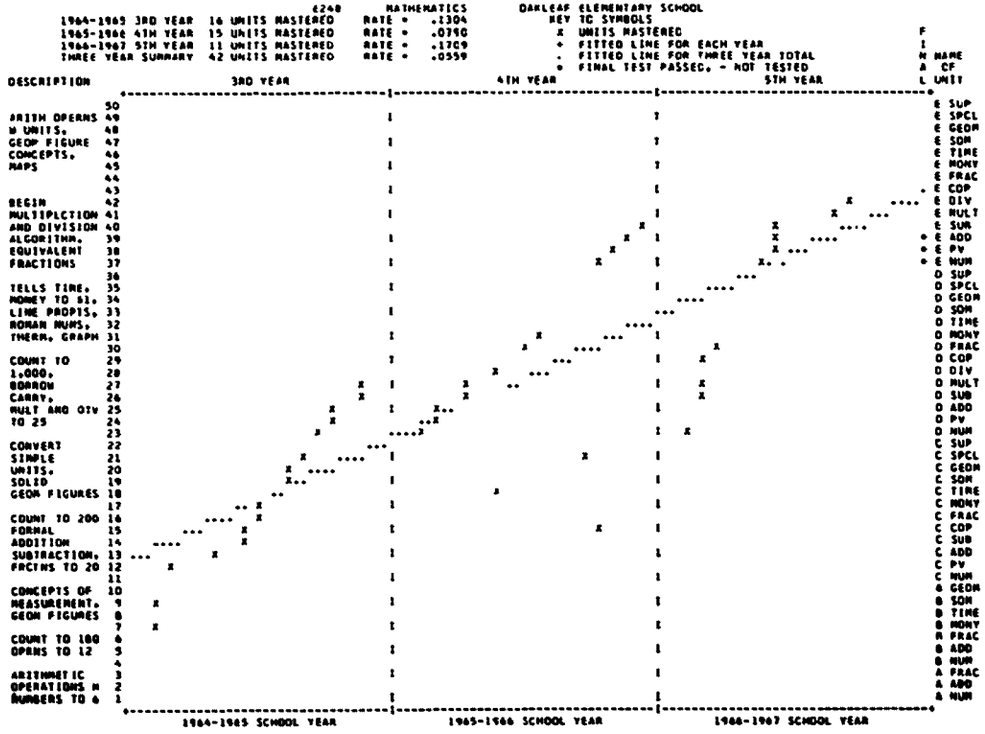


Figure 12 *Charlene*

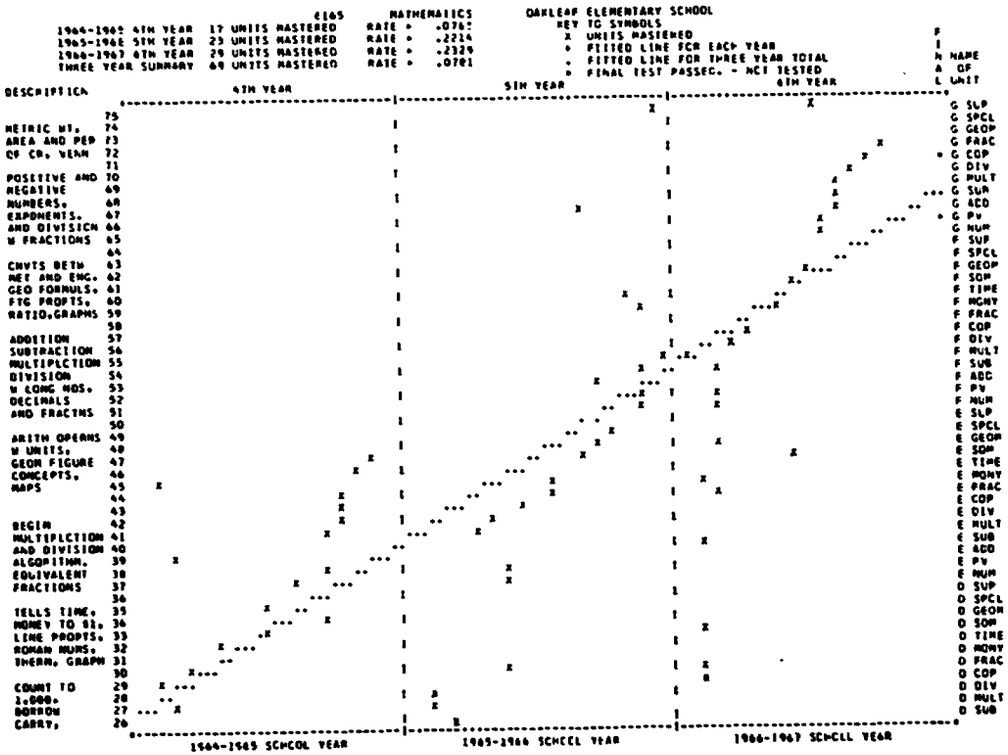
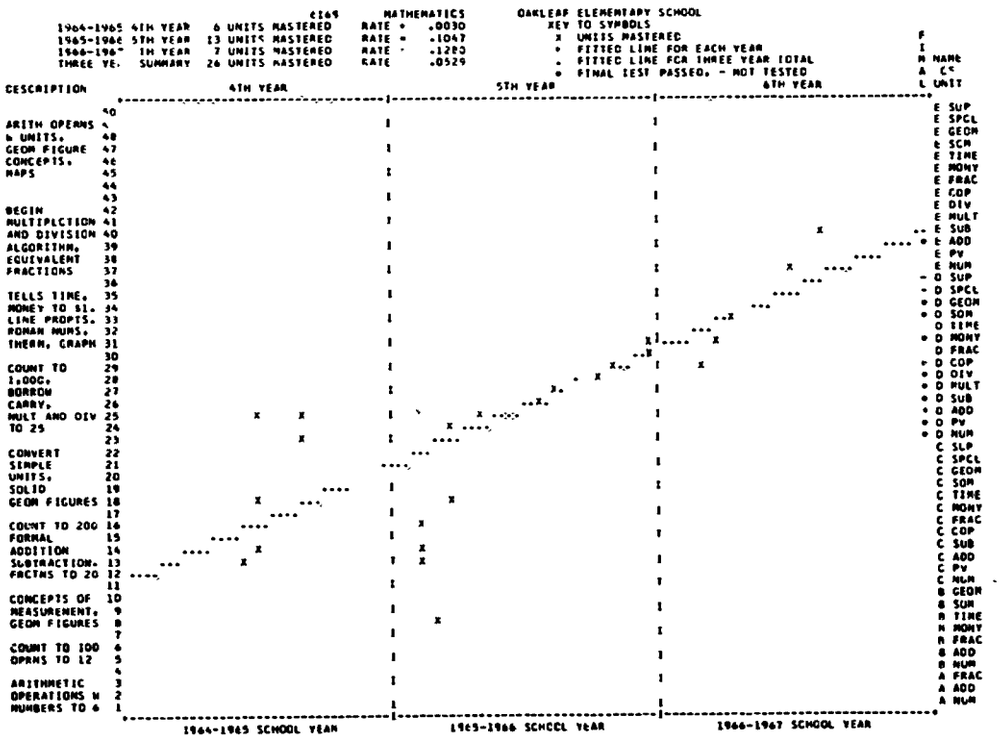


Figure 13 *Diane*



It is of further interest to examine the distribution of attained achievement level in mathematics as the students, teachers, and researchers obtain practice in individualizing instruction from year to year. The bar charts in Figure 14 show each grade in Oakleaf for the years of 1965, 1966, and 1967. Figure 15 shows the total school over these three years. The height of a bar indicates the number of students ending up the school year at a particular level of the mathematics curriculum. On the horizontal axis, each level is divided into two parts; level A, the kindergarten level, is not shown. Figure 14 shows that in grade 6, approximately eight students were working at level G_1 at the end of the 1966-67 school year; at the end of the 1965-66 school year, only one sixth grader had reached level G_1. The general trend in Figure 14 is that attainment levels over the three years are moving up the mathematics continuum. Figure 15 shows this again and also shows that the spread of attainment is larger in 1967 than in the previous years. Notice in Figure 14 that in the first year of individually prescribed instruction in the first grade, the system did not provide effective means for allowing the children to move on their own, and they ended up as a group in various units in the first half of level B. As a result of this experience, level A was moved into the kindergarten, and materials and procedures were revised so that preparation of students for individualized learning could take place at that time, permitting them to be more self-directive in the first grade.

Conclusion

In conclusion, it appears that the Mark I phase of individually prescribed instruction provides a start toward individualizing certain aspects of a school curriculum—a start that can be specifically studied, revised, and improved. We are encouraged now to intensively investigate the relationships and the conditions of learning that underlie the attainment of a broad spectrum of educational goals. It is conceivable that individualized instruction will find its major value in attaining not only achievement objectives but other educational goals such as self-direction, self-initiation of one's learning, and the feeling of control over one's learning environment. Success in reaching these outcomes of learning is difficult to measure, but we plan to try to do so.

The thesis suggested at the beginning of this paper should be restated. It is that progress in the design of systems for the individualization of instruction can only be seriously considered if we try to

Figure 14

Curriculum Levels Attained by Students in Each Grade in June of 1965, 1966, and 1967

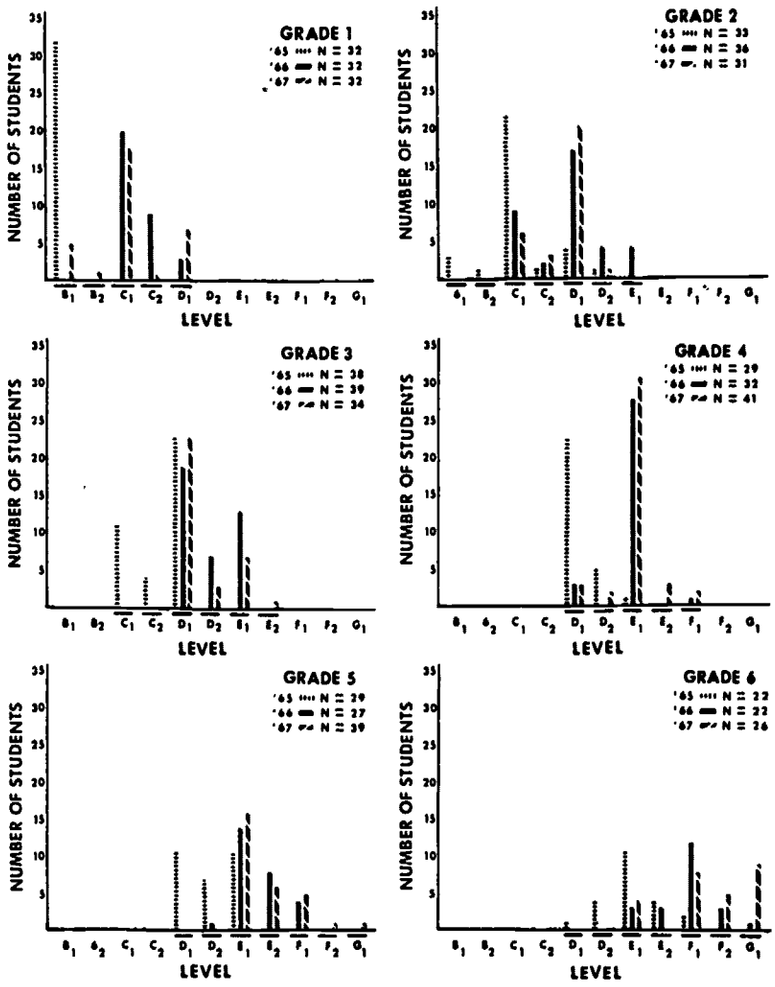
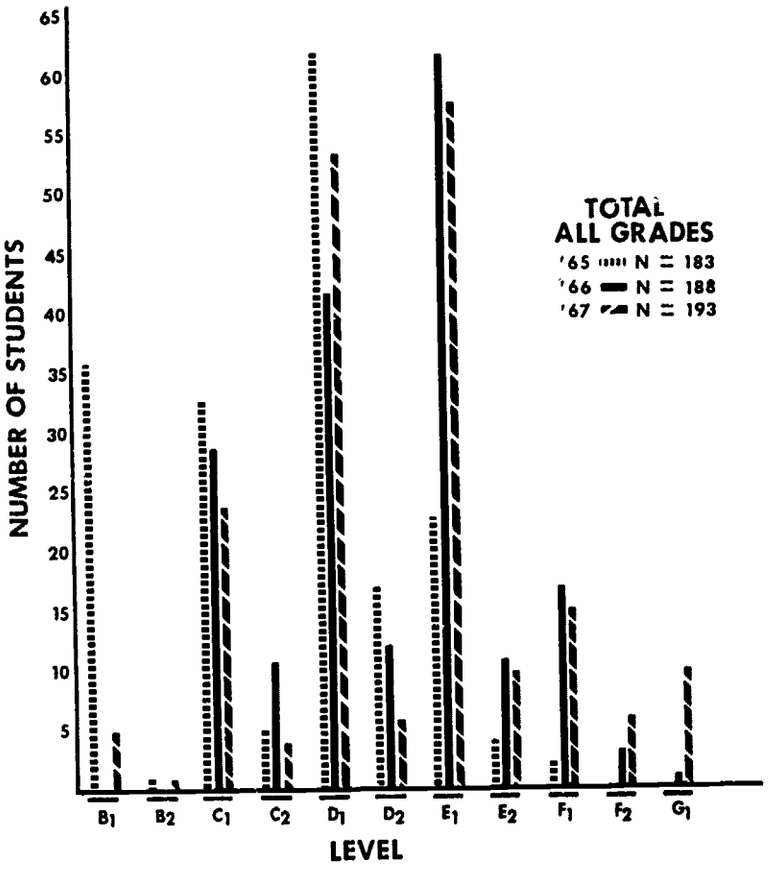


Figure 15

Curriculum Levels Attained by Students in Each Grade in June of 1965, 1966, 1967



understand the relationships that underlie the practices we implement. I suspect that neither we in our work nor anyone else will have any significant or reliable success with a program of individualized instruction unless such understanding is obtained by sustained research and experimentation in the schools. As I have indicated, the first step is for the schools to allow extensive redesign of their curriculum and instructional procedures, and once this takes place, to be disabused of the notion of "instant progress through innovation," and to permit sustained study of the variables and effects involved. Fortunately, this is occurring in a number of places. For example, in our work with Research for Better Schools (the Regional Laboratory in Philadelphia sponsored by the Office of Education), 23 elementary schools in various parts of the country will be obtaining data for the analysis and improvement of the individually prescribed instruction procedure. As we begin to study these data and the even more detailed data obtained from computer management of individualized instruction in particular schools, the hope is that we will be able to uncover further knowledge about the relationships between individual differences and instructional methods. With this knowledge in hand, we should be able to bring the aspirations of the authors of the 1925 NSSE Yearbook to a happier ending.

REFERENCES

1. Bialek, H. Personal communication.
2. Cronbach, L. J. and Gleser, G. C. *Psychological tests and personnel decisions.* Urbana, Illinois: University of Illinois Press, 1965.
3. Fleishman, E. A. The description and prediction of perceptual-motor skill learning. In R. Glaser (Ed.), *Training research and education.* New York: John Wiley & Sons, 1965. Pp. 137-176.
4. Gage, N. L. and Unruh, W. R. Theoretical formulations for research on teaching. *Review of Educational Research,* 1967, *37,* 358-370.
5. Washburne, C. N. Adapting the schools to individual differences. *Yearbook of the National Society for the Study of Education,* 1925, 24, Part II.

An Evaluation Model for Professional Education—Medical Education

CHRISTINE H. MCGUIRE
College of Medicine, University of Illinois

To a generation conditioned to think of the physician as a member of a rather conservative profession dedicated to the defense of its own nostalgic image of nineteenth century individualism, any comparison between medical education of today and progressive education of the thirties may appear to be the product of a completely private fantasy. But to an observer of both, there are some striking similarities between the two with respect to the motivation for change, the values to be sought in change, and the zest with which change is pursued.

The desire to modify the educational system has been stimulated in the medical educator of today, as it was in the secondary school educator of the thirties, by social forces that, if ignored, threaten to produce serious disorder. Three such forces are of special importance in medicine: 1. the competition for young talent from the new and glamorous fields of nuclear physics and the various hyphenated bio-sciences that has caused some worried head-shaking about whether the same caliber of bright and dedicated young men and women are now being recruited into the field as were easily won to it a generation ago; 2. the so-called explosion of knowledge to which some have reacted with a strident demand to lengthen the period of professional training (a period that already extends well beyond age 30 for many aspiring physicians) and to which others have responded by an iconoclastic appeal to minimize the emphasis on "facts" and to eliminate "unnecessary" requirements from the curriculum; and 3. the manpower shortage created in part by unrelenting public demand for more extensive medical services rendered in unfamiliar institutional settings.

The discussion and experimentation in medical education in response to these forces are not unlike the ferment in secondary education

produced by analogous social forces of the thirties. But the similarity between the two movements goes beyond that of a common type of motivation; they are also alike in their stress on attitudes, values, and ethical systems—in short, in their relative emphasis on the importance of educational influences and outcomes in the affective domain. Finally, the enthusiasm—almost missionary zeal—which characterizes, and the controversy which surrounds, the innovators of this period are strangely reminiscent of the emotional fervor invoked and evoked by the progressive educators of that earlier era.

It is from this highly personal orientation that I wish to develop the thesis that a climate conducive to change has developed in the world of medical education and that, in this setting, a new approach to evaluation is beginning to make a significant contribution to the systematic modification of medical education. Moreover, the role of evaluation in the reform of professional education is, in some ways, analogous to that which a different evaluation model played in the earlier transformation of general education. To support this thesis, I should like to present three case descriptions that illustrate some of the ways in which evaluation is currently being incorporated as an integral part of a dynamic system of professional education. All three are drawn from the arena of medical education: The first, which will be discussed more fully than the other two, is an institutional mechanism for systematic data collection and regular multi-channel feedback that has been deliberately established by the faculty of one college of medicine. The second is a research study now under way at the level of graduate medical education and specialty board certification. The third represents a general schema for on-going institutional self-study that has been made operational to varying degrees in some half-dozen medical schools.

An Institutional Mechanism for Systematic Data Collection and Regular Multi-channel Feedback

In 1959, at the official request of a standing committee of the faculty, the newly established Office of Research in Medical Education of the University of Illinois College of Medicine made a series of studies of the "climate for learning" at that institution. Some of the findings of those studies supported the view that the then current system of student examination, grading, and promotion not only failed to provide evidence regarding student achievement with respect to many of the most important goals of medical education, but that it actually jeopardized their

attainment by exacerbating tendencies toward fragmentation of learning, by focusing student attention on esoteric or trivial detail, and by intensifying unhealthy competition among students for grades and among departments for students' time and attention.* To remedy this situation and to implement the principle that setting of standards for certification is an institutional, not merely a departmental, function, the responsible faculty committee recommended that a college-wide system of comprehensive examinations be established on an experimental basis. It further recommended that the task of constructing, administering, and reporting all instruments used for purposes of official student assessment be assigned to a specially created faculty Committee on Student Appraisal.

During the first year of its existence, that Committee, unwilling to rely simply on purely pro forma statements of course objectives, not only reviewed official documents but also initiated extended discussion with representatives of the several departments in an effort to develop a comprehensive and coherent set of institutional goals and standards in terms of which student progress could be assessed. The resulting behavioral objectives were then categorized into the familiar cognitive, affective, and skills domains, and appropriate techniques for assaying each were explored. It was decided to rely on comprehensive written examinations to measure the cognitive goals and some of the skills and to use practical laboratory and clinical examinations to measure other of the psychomotor skills. Habits and attitudes were to be assessed by systematic accumulation of anecdotal records in whatever quantity individual faculty members felt able and willing to supply them.

Since it serves to illustrate the total process, I shall limit the discussion here to a description of the procedures evolved for assessment in the cognitive domain only. In this area, the Committee decided that an integrated examination should be constructed for each year of the four-year medical curriculum, and that the comprehensive examinations developed for administration at the end of the second, third and fourth years should sample—to some degree, at least—the content and intellectual skills toward which the curriculum of all *previous* years had been directed. I stress this policy not only because it has facilitated collection of longitudinal data but also because it has clarified (some

*For example, analysis of grades revealed that by the end of the first year, students in the highest 10 percent of the class on the Watson-Glazer test of critical thinking had fewer honor grades and more failing grades than those in the lowest 10 percent of the class (12).

might say sharpened) fundamental issues with regard to the sequence and integration of the medical curriculum, and has been instrumental in conveying to students the value the faculty places on continued application of facts and concepts explicitly treated only in the first year or two of the curriculum. Since, for practical reasons, it is necessary to assign responsibility for the development of each comprehensive to a different subcommittee, the membership of each subcommittee is arranged to include faculty from every level of the medical curriculum in order to facilitate implementation of this principle of longitudinal integration.

These subcommittees meet with a designated representative (liaison examiner) from every department that offers required courses at the relevant curricular level. Together the representatives and the subcommittee develop specifications with respect to both content and behavior for the comprehensive under consideration, discuss suggestions for the types of exercises that will meet these specifications, and assign responsibility for the initial preparation of such exercises.

In order to meet the specifications established by the subcommittees, it has been necessary to extend present test techniques in new directions. For example, in developing exercises designed to measure ability to interpret data, it is necessary to take account of the fact that data in medicine come in exceedingly varied forms. They may be in the form of gross or microscopic specimens obtained at autopsy; they may be tracings on an oscilloscope; they may be sounds heard through a stethoscope; they may be numbers on a laboratory report or films from X-ray; they may be the appearance or movements of a patient, or his responses to various inquiries or maneuvers. Thus, in order to develop valid, as well as objective and reliable, assessments, the Appraisal Committee has constructed interpretation-of-data exercises based on video tapes of patient interviews, color movies of a patient examination or an autopsy (3), high fidelity tapes of heart, breath, lung and abdominal sounds suitable for replay through individual stethophones, and standard photographic reproductions of all manner of visual findings such as X-rays, lesions, eye grounds and biopsy specimens. Exercises based on such materials require the student to demonstrate that he can make accurate observations of the data presented, can see their significance and possible interrelations, can recognize basic biochemical or pathophysiologic processes that would explain them, can anticipate other findings that might be associated with them, or drugs that might produce or reverse them—in short, that he can take a multi-disciplinary

approach to the interpretation of relevant data presented in as nearly realistic a form as possible.

Similarly, building on the idea of programmed examinations which Dr. John Hubbard of the National Board of Medical Examiners described to you at an earlier conference (4), the Appraisal Committee has constructed a number of simulation exercises to test the complex skills of gathering data and making judgments. These simulations are in written form and are thus amenable to group administration and computer scoring (9). In effect, they constitute branched problems in patient management or in laboratory investigation that require sequential analysis and decision. A clinical simulation, for example, is initiated by a brief verbal description of the patient's chief complaint or by a short color film in which the patient describes his current illness. The examinee must then decide how he will first approach this patient —*i.e.*, what, if any, further work-up seems indicated at this point. He records this decision by erasing the opaque overlay on a specially constructed answer sheet and finds an instruction directing him to the section designated by his choice. Here he is confronted by a long list of possible interventions that will yield further information about the patient. He may select as many or as few procedures as seem appropriate in light of the specific circumstances obtaining at this stage in the problem. He again records each choice by erasing the appropriate overlay to find the results of that intervention presented in realistic verbal, visual, or auditory form resembling that which the physician is accustomed to encountering.

On the basis of these new data he must decide upon the next step he wishes to take. Each such problem is constructed to allow both for different medical approaches and for variation in patient responses appropriate to these several approaches. Even the complications which must be managed differ from student to student depending (as they do in the office or clinic) on the unique configuration of prior decisions each student has made. For some, the erasures will reveal an instruction to skip one or more sections of a problem because the approach they have chosen is effective in avoiding potential complications with which others must cope. If, however, at any stage the examinee orders something harmful or fails to take measures essential to the recovery of the patient, he uncovers a description of the clinical features of the complication that has developed. He is then directed to a special section where he has the opportunity to take heroic measures to rectify his previous errors; if the remedial measures are inadequate he may be

instructed that the problem is terminated because the patient has suffered a relapse and has been sent to another hospital, or has been referred to a consultant, or has died.

After the new exercises have been prepared and revised until they are acceptable to other experts in the author's specialty, they are subjected to detailed critical review by the subcommittee which, as noted above, includes representatives from both basic science and clinical disciplines. This system of developing and reviewing exercises has three obvious advantages: First, it imposes some checks and balances on the specialist that assist him in focusing on the basic concepts and cognitive skills for which every physician, irrespective of his future specialty, should be held accountable. Second, it requires experts in different specialties to explore and reconcile possible differences in approach to specific, common problems and thus facilitates a comprehensive and integrative consideration of such problems. Third, since the comprehensive for all four years of the medical school are, in the last analysis, under the aegis of a single parent committee which operates on the principle that the student is responsible not only for the current year's work but also for all that has preceded, the system provides simple machinery for gathering data on the extent to which understanding of fundamental patho-physiologic principles gained during the basic science component of the curriculum is augmented or diminished during the subsequent clinical component and, conversely, on the extent to which students are already able to make accurate clinical applications of these principles prior to any clinical study.

Once an exercise has been accepted for inclusion in the pool of examination materials, it is coded according to both the content and the cognitive process which it purportedly samples. This procedure is followed in order to assist in compiling an examination that conforms as closely as possible to the specifications initially set for it and to facilitate diagnostic scoring of subtests defined according to both content and behavioral categories.

The final step in the preparation of each examination consists in determining the minimum acceptable performance on the total comprehensive and on each subtest *prior to its administration.* This is accomplished by a procedure similar to that originally described by Nedelsky (11). According to this method, each item is assigned a "Minimum Passing Level" which represents, in effect, an expert estimate of the chances that a barely passing student would have of selecting the best answer to that question. Estimates for all items are com-

bined in a manner to yield a "Minimum Passing Level" for a total test and for each subtest. Since this procedure for defining the passing score constitutes the application of predetermined, absolute standards of competence that are independent of the actual performance of any group, the failure rate in any class can, in principle, vary from zero to 100 percent. This fact has important implications for the nature of the feedback that can be furnished faculty and students. For example, in our experience the failure rate on various subtests has ranged from zero to well over 50 percent. Such fluctuations have been used not only to flag possible trouble spots in the program but also to help identify the real improvements in student achievement that are so often obscured by more common procedures for setting standards.

In order to provide the variety of data needed by different groups, all examinations are scored by a computer program which, in addition to individual total and diagnostic subscores, yields unusually complete test and item statistics (6). Before the results of any examination are officially reported, the several contributors to the comprehensive and the subcommittee responsible for having constructed it review the item data with a view to identifying and eliminating any defective questions. Following this review, the entire examination is re-scored, and a report is prepared summarizing the main characteristics of the examination and its results and embodying detailed information about the pre-established minimum passing levels and the number of students who fail to achieve each. This summary, together with a set of individual reports indicating each student's performance on the comprehensive examination and on the official skills examinations, and his ratings with respect to various professional habits and attitudes, is transmitted to the Promotions Committee. This latter committee, on the basis of the data supplied to it by the Appraisal Committee, has the responsibility for making all official decisions with regard to the requirements to be met by any student who is deficient in any respect. In making these decisions as to whether the student will be promoted, required to repeat an examination or an entire year, or dropped, the Promotions Committee has available an unusual amount of information not only about the different dimensions of student achievement but also about the reliability and validity of the various measures to assess this achievement. Students, their instructors, and their academic advisers are provided with a description of the examination, the individual reports (including all diagnostic scores), and detailed information on the preestablished standards of satisfactory performance as well as on the actual

performance of the total group. Finally, in addition to the above-listed information, departments and standing committees of the faculty are provided with detailed item analysis and subtest data to assist them in identifying major strengths and weaknesses in the programs for which they are responsible. In these reports to departments and committees, special attention is called to two sets of results: 1. the performance of students at different levels of the curriculum on identical exercises that have been incorporated in several comprehensives, in order to assess student progress toward goals that are common to the entire four-year curriculum; and 2. any unusual trends in numbers of students failing to meet the preestablished standards. Dialogues between the Appraisal Committee and the relevant departments are initiated regarding possible hypotheses to explain these trends. These dialogues will involve such questions as: Are the preestablished standards really appropriate for the non-specialist? Is the test, in fact, a reliable and valid sample of relevant student achievement? Are there changes, either planned or inadvertent, in the instructional program that may help to explain the results? Are there general changes in student motivation or environmental press that may have affected student achievement?

To what extent are the data provided being exploited as a basis for decision making and to what extent, if any, has this systematic accumulation of data resulted in important educational changes within the institution? A considerable body of anecdotal information is now available which strongly suggests that the feedback to students has helped them to clarify the important goals of the instructional program, and that the character of the examinations has helped students to focus their attention on learning to *apply* the vast body of knowledge they are acquiring. Further, there is considerable evidence that many students have found the diagnostic information provided them increasingly useful in directing their further study. Finally, by relieving the instructor of the necessity of being both mentor and judge and by creating a situation in which a student's performance is compared not to that of his colleagues but to a preestablished standard which theoretically all can meet, the system has—in principle at least—removed some of the impediments to a more mature and responsible relation among students and between students and their instructors.

Any attempt to claim a direct relation between this new system of student evaluation and the changes that are occurring in curricular and instructional programs at the University of Illinois would quite properly be met with considerable skepticism. In this connection, therefore, let

me state merely that the present system of student appraisal has made it possible to collect exceedingly useful longitudinal and cross-sectional data on the consequences of these changes. Moreover, the policy committees responsible for modifying the program are not only turning to the data "after the fact" but are increasingly incorporating provision for systematic evaluation as an essential component in proposals for curricular reform. Finally, the effects of the appraisal system on instructional methodology are particularly subtle and exceedingly diverse. Individual instructors and departments vary greatly in the degree to which they exploit the data provided them. Some have requested assistance in investigating a very specific hypothesis about experimental modifications in instructional methodology; others report changes in their general approach to teaching and learning; some have not been aware of the information available, and others have expressed no interest in it.

In the short paragraphs above, I have described a specific situation in which student evaluation is an integral part of a program that automatically provides reciprocally corrective feedback and that has the potential of becoming a smoothly functioning cybernetic system. The second case, which I shall treat much more briefly, illustrates a quite different use of the same type of evaluation model—namely, as part of a research system.

Research on Assessment of Professional Competence

Some four years ago, in cooperation with a major surgical specialty board, the Center for the Study of Medical Education undertook a joint study of certifying procedures used in assessing professional competence in that specialty. The investigators felt that the availability of valid and reliable measures of the various aspects of competence was a prerequisite for scientific development of more efficient and effective training programs, which, in turn, could make a substantial contribution to the more rational utilization of scarce manpower resources.

The first stage of the research was devoted to a determination of the essential components of competence in the specialty under study; the second, to an investigation of the adequacy of current certification techniques as measures of these components; and the third, to development of instruments that would yield more relevant, valid, reliable, and comprehensive assessments.

During the first phase of the investigation, a critical incident study (2) was undertaken to provide an empirical basis for a behavioral descrip-

tion of the essential components of professional competence. Over 1,700 incidents were collected from more than 1,000 specialists representing various age groups, types of affiliation, and subspecialty interests. An empirical classification defining 94 critical performance requirements, grouped into 9 major categories of competence, was derived from the incidents (10). That operational definition of the essential components of competence has since been employed to direct all subsequent stages of the study and has served as the basis for the development of a blueprint specifying the content and skills to be sampled in the certification process.

In the second phase of the joint study, a systematic process analysis (7) was made of both the written and oral examinations currently used by the specialty board for certification purposes. The written examination was analyzed by three subject-matter experts, and each item was classified according to the intellectual process required in responding to it. In the final classification, over 50 percent of the items were unanimously rated by all experts as measuring predominantly recall and recognition of isolated information; fewer than 25 percent of the questions were thought by any expert to involve even simple elements of interpretation of data or familiar problem solving.

Study of the oral examinations was conducted by a team of five physicians and three educators trained in systematic observational analysis, using a specially developed, pretested form for the descriptive recording of the observations. The team made 158 observations of 144 individual half-hour examinations, which represented a stratified sample of the more than 2,000 individual examinations administered to candidates applying for specialty board certification in January 1965. Although it is unlikely that observers were able to record each question put to a candidate, 6,868 were recorded, and each was classified according to the nature of the intellectual process it seemed to elicit. Analysis of the observations (8) indicated (a) that these oral examinations measured predominantly the candidate's ability to recall (rapidly and under stress) isolated fragments of information (some 70 percent of the examiner-candidate exchanges were rated at this level); (b) that candidates only rarely (in fewer than 2 percent of the exchanges) cited evidence for their answers; and (c) that standards employed in judging performance were not always clear nor were they uniformly applied.

Findings from a subsequent factor analysis provided further support for the conclusions based on this initial process analysis—namely, that the traditional examinations, both oral and written, measured predomi-

nantly a candidate's ability to recall specific information and that other performance requirements identified in the critical incident study were not being assessed.

The third stage of the study was therefore devoted to the development and analysis of certain experimental techniques designed to measure previously neglected aspects of professional competence. In order to assess the more complex cognitive skills, modified types of multiple-choice questions and written simulations of patient management problems as described above were developed. In addition, in order to assess those important skills and attitudes in dealing with patients and colleagues that are not readily assayed by more conventional techniques, a number of oral simulations and role-playing situations were introduced into the certifying examination. Three of these, designed by Mr. Harold Levine of the Center staff (5), may be of special interest. The first, designed to measure the ability to gather information, to solve diagnostic problems, and to communicate with patients, consists of a 20-minute simulated diagnostic interview in which the examinee assumes the role of a physician and elicits diagnostic information from the examiner, who is cast in the role of the patient. The second, designed to measure the ability to communicate with and relate effectively to patients, consists of a 10-minute simulated proposed treatment interview in which the examinee assumes the role of a physician who must explain a proposed treatment to an examiner who is cast in the role of a specified patient whose cooperation must be won. The third, designed to measure the ability to communicate with and relate effectively to colleagues, consists of a 30-minute simulated staff conference in which five candidates discuss the management of two standardized cases. Each of the role-playing exercises is scored on a standardized rating scale on which the specific behavioral factors to be rated and the criteria for assigning ratings are described in detail.

Finally, both for purposes of validating other techniques and of obtaining data on certain skills, habits, and attitudes that cannot be assessed in the structured examination, training chiefs who have supervised examinees during their training are asked to rate them on a standardized 12-point scale with respect to each of the following factors: (a) ability to recall information; (b) ability to use information to solve problems (inductive and deductive reasoning); (c) ability to gather information; (d) clinical judgment (tendency to take all important criteria into account in deciding on treatment and weighting the criteria appropriately); (e) surgical skill; (f) ability to relate to

patients; (g) ability to relate to colleagues; (h) demonstrating appropriate moral and ethical standards; and (i) overall competence required of a physician.

Studies of both sampling and inter-rater reliability were conducted on all traditional and experimental techniques, and modifications in the design and scoring of all instruments were made in light of the results of these studies. In addition, content, construct, and concurrent validity of the various examinations have been investigated by a variety of techniques. For example, in the analysis of construct validity, studies of the relation between performance and level of training reveal that performance on tests designed to measure general information in the discipline or to assess decisiveness about treatment are most highly correlated with level of training; those designed to assess thoroughness of diagnostic work-up or ability to relate to patients are least so. These results are consistent with other information about the nature and relative emphases of most training programs. Further, studies of the influence of experience and type of practice on responses to the written simulations of patient management problems reveal the same relationships as described in earlier observational studies of practitioner performance (1, 13). For example, among certified specialists in practice, performance on the written simulations is negatively correlated with age and positively correlated with closeness of academic affiliation.

Finally, concurrent validity of the various measures was investigated through correlational studies of the relationship between supervisors' ratings and performance on the various oral and written tests. Care must be taken in interpreting the results of these studies since there was undoubtedly a certain amount of variance attributable to error in both the ratings and the test scores. Furthermore, the tests did not assess some of the factors included in the ratings (*e.g.*, surgical skill and ethics); and alternatively, supervisors may have been unable or unwilling to observe some of the important behaviors assessed by the tests. With these reservations, the following tentative generalizations seem reasonable in light of available data: Supervisors take many factors into account in evaluating the overall competence of their residents. The most important of these are what the supervisors identify as "problem-solving ability," "clinical judgment," and "ability to relate to colleagues." The best predictor of what they mean by "problem-solving ability" is the score on one of the conventional oral examinations. The best predictor of what they refer to as "clinical judgment" is the multiple-choice examination (though the score on the treatment component

of the written simulations also makes a significant contribution to the prediction of this performance factor). The best predictor of "ability to relate to colleagues" is the score on the simulated-patient interview involving explanation of a proposed treatment to a simulated patient. In light of the findings from the factor analytic studies noted above, these results suggest that supervisors' ratings of "problem-solving ability" and "clinical judgment" are heavily influenced by a common factor of "general information."

Studies such as those described above not only have contributed to an analysis and validation of the experimental certifying techniques, but also have directly or indirectly influenced the introduction of the following modifications in training and certifying practices in the surgical specialty under study:

1. A regular procedure has been instituted for the annual collection of background data, ratings of program directors, and examination performance data on each resident in each year of training. These are to be used in advising residents and in analyzing the variables related to different patterns of competence and differential rates of achievement.

2. In accord with the components of competence defined in the critical incident study, this specialty board has established an examination blueprint specifying the cognitive processes to be evaluated and the subject-matter content to be sampled, and the weight to be assigned each in the certifying examination, and has developed procedures for assuring that these specifications are met.

3. The total certifying process is being redesigned to yield evidence on a number of aspects of professional competence not previously assessed and to assure greater standardization in the measurement of others.

4. The scoring and reporting system is being redesigned to yield a profile of performance in which evidence from several sources will be combined to produce the most reliable assessment of each factor; certification will be made on the basis of predetermined standards of excellence with respect to these factors, irrespective of their effects on the failure rate.

5. A program for training of examiners in the development of standardized materials and in the administration and scoring of oral examinations has been instituted.

6. A four-year extension of the present study is projected, which will have as its primary focus controlled experimentation in curriculum and instructional techniques employed in graduate education in this specialty. During this phase of the study, experimental modifications will be introduced into selected training programs, and the consequences of their effects on resident achievement will be assessed through cross-sectional and longitudinal studies as a means of determining the relationship among in-put, training, and out-put variables.
7. Finally, a ten-year follow-up is now being planned to investigate both the predictive validity of the new certifying procedures and the long-term effects of modifications in the training programs.

This, then, is a very brief overview of a specific research study which has already had a direct impact on training and certification procedures at the graduate level in at least one surgical specialty and has been viewed with considerable interest as a model for systematic self-study by other professional groups.

An Evaluation Model for Self-study

The third case—a unique kind of institutional self-study—represents a somewhat different type of evaluation model. Fortunately, I need merely allude to it since it has been fully described in an excellent monograph by Dr. Paul Sanazaro (14), recently published by the Association of American Medical Colleges. In the four institutions in which this systematic self-study has been made fully operational, virtually the entire faculty has been involved in the continuing collection and provision of comprehensive data on in-put, environmental, and out-put variables, and their interaction in the local institution. These data are considered in an ongoing series of special faculty seminars where, with the aid of educational consultants who have assisted in the original research design, they are put in the context, on the one hand, of educational theory about learning, curriculum, evaluation, and the like, and, on the other, of comparative data from various national, regional, and peer groups of institutions. The institutional data and contextual materials are further utilized in a subsequent series of executive sessions of the faculty where they serve as a basis for policy decisions. At this time, provision is generally made for evaluation of program modifications and their continued monitoring through systematic data collection.

Summary Comment

It has not been my purpose to outline a rigid evaluation model, but rather to indicate some of the types of situations in which a particular approach to educational evaluation is currently being employed at many different levels of medical education. Increasingly, the products of medical education are being studied by systematic evaluation procedures which include: empirical determination of essential components of professional competence, employment of simulation techniques to supplement more conventional methods of assessment, application of preestablished standards, and utilization of numerous feedback mechanisms to assure fuller exploitation of evaluation data. Such evaluation studies are being employed not only to assess individual achievement of critical performance requirements, but also to identify differential rates and patterns of progress toward these goals, to determine the relation between these patterns and important independent variables in the learning situation, to guide curricular development, and to provide evidence of value in re-defining the goals themselves. It seems clear that systematic feedback from this type of evaluation process has been of value in a consideration of problems of professional education, irrespective of the point in the planning cycle at which the demand for change may arise.

REFERENCES

1. Clute, K. F. *The general practitioner: a study of medical education and practice in Ontario and Nova Scotia.* Toronto: University of Toronto Press, 1963.

2. Flanagan, J. D. The critical incident technique. *Psychological Bulletin,* 1954, 51, 327-358.

3. Hubbard, J. P., *et al.* An objective evaluation of clinical competence. *New England Journal of Medicine,* 1965, 272, 1321-1328.

4. Hubbard, J. P. Programmed testing in the examinations of the National Board of Medical Examiners. In Anne Anastasi (Ed.), *Testing Problems in Perspective.* Washington, D. C.: American Council on Education, 1966. Pp. 195-207.

5. Levine, H. G., and McGuire, C. H. Role-playing as an evaluative technique. *Journal of Educational Measurement* (on press).

6. Lewy, A., and Crawford, W. Scoring test battery: a program for the IBM 7094. *Educational and Psychological Measurement*, 1966, 26, 1, 185-188.

7. McGuire, C. H. A process approach to the construction and analysis of medical examinations. *Journal of Medical Education*, 1963, 38, 556-563.

8. McGuire, C. H. The oral examination as an assessment of professional competence. *Journal of Medical Education*, 1966, 41, 267-274.

9. McGuire, C. H., and Babbott, D. Simulation technique in the measurement of problem-solving skills. *Journal of Educational Measurement*, 1967, 4, 1-10.

10. Miller, G. E., *et al.* The orthopaedic training study—a progress report. *Bulletin of the Academy of Orthopaedic Surgery*, 1965, 13, 8-11.

11. Nedelsky, Leo. Absolute grading standards for objective tests. *Educational and Psychological Measurement*, Spring, 1954, 14, 3-19.

12. Office of Research in Medical Education. *The relation of academic achievement to skill in reading, critical thinking and three-dimensional perception.* Annual Report to the Faculty. Chicago: University of Illinois Medical Center, September 1961. 64 pp.

13. Peterson, O. L., *et al.* An analytical study of North Carolina general practice. *Journal of Medical Education*, 1956, Part 2, 31, 1-165.

14. Sanazaro, P. *Educational self-study by schools of medicine.* Evanston, Illinois: Association of American Medical Colleges, 1967.

Session II

Theme:
New Approaches
to Instruction

Computer-based Instruction in Initial Reading

RICHARD C. ATKINSON
Stanford University

Almost four years ago, Patrick Suppes and I initiated a project under a grant from the Office of Education and the Carnegie Foundation of New York to develop and implement a program of computer-assisted instruction (CAI) in initial reading and mathematics. Because of our research interests, Suppes has taken responsibility for the mathematics curriculum, and I have been responsible for the initial reading program.

At the beginning of the project, two hurdles had to be overcome. There was no lesson material in either mathematics or reading suitable for CAI, and an integrated CAI system had not yet been designed and produced. The development of the curricula and the development of the system have been carried out as a parallel effort over the last four years with each having a decided influence on the other.

In this paper, I would like to report on the progress of the reading program, with particular reference to the past school year when for the first time a sizable group of children received a major portion of their daily reading instruction under computer control. The first year's operation must be considered essentially as an extended debugging of both the computer system and the curriculum materials. Nevertheless, some interesting comments can be made regarding both the feasibility of CAI and the impact of such instruction on the overall learning process.

Before describing the Stanford Project, a few general remarks may help place it in proper perspective. There are, first of all, three levels of CAI. Discrimination between these three levels is based not on hardware considerations but principally on the complexity and sophistication of the student-system interaction. While an advanced student-system interaction may be achieved with a simple teletype terminal, the most primitive interaction may require highly sophisticated

computer programs and elaborate student terminal devices.

At the simplest interactional level are those systems that present a fixed, linear sequence of problems. Student errors may be corrected in a variety of ways, but no real-time decisions are made for modifying the flow of instructional material as a function of the student's response history. Such systems have been termed "drill-and-practice" systems and at Stanford University are exemplified by a series of fourth-, fifth-, and sixth-grade programs in arithmetic and language arts designed to supplement classroom instruction. These particular programs are being used in several different areas of California and also in Kentucky and Mississippi, all under control of one central computer located at Stanford University. Currently as many as 2,000 students are being run per day; it requires little imagination to see how such a system can be extended to cover the entire country (5, 9, 10).

At the most complex level of student-system interactions are "dialogue" programs. Such programs are under investigation at several universities, but to date, progress has been extremely limited. The goal of the dialogue approach is to provide the richest possible student-system interaction in which the student is free to construct natural-language responses, ask questions in an unrestricted mode, and in general, exercise almost complete control over the sequence of learning events.

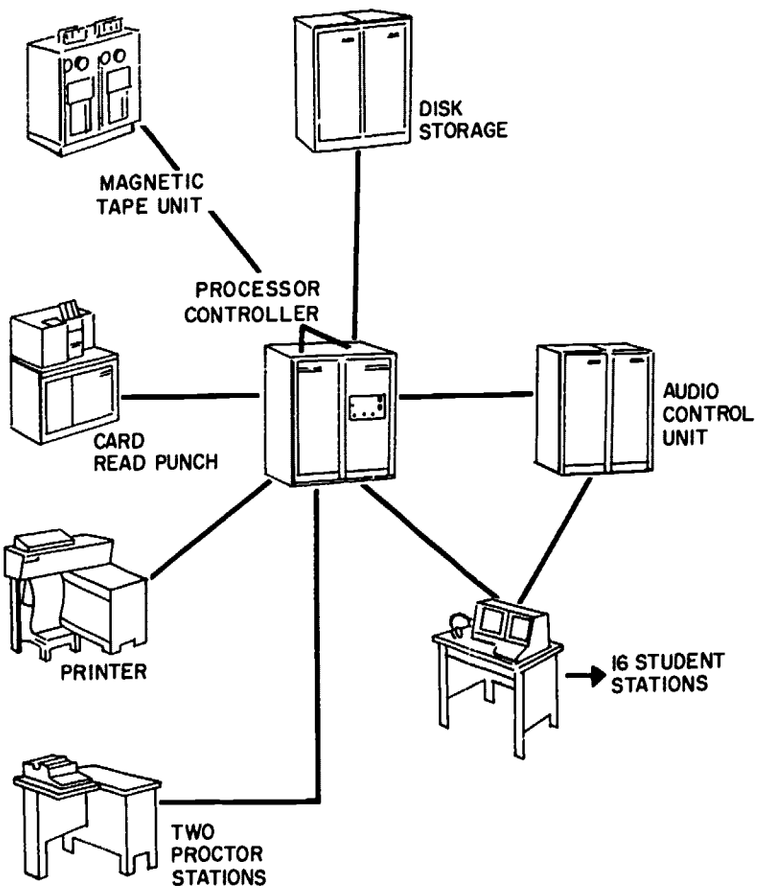
"Tutorial" programs lie somewhere between the above extremes of student-system interaction. Tutorial programs have the capability for real-time decision making and instructional branching contingent on a single response or on some subset of the student's response history. Such programs allow students to follow separate and diverse paths through the curriculum based on their particular performance records. The probability is high in a tutorial program that no two students will encounter exactly the same sequence of lesson materials. However, student responses are greatly restricted since they must be chosen from a prescribed set of alternatives, or constructed in such a manner that a relatively simple text analysis will be sufficient for their evaluation. The CAI Reading Program is tutorial in nature, and it is this level of student-system interaction that I want to discuss.

The Stanford Tutorial System

The Stanford Tutorial System was developed under a contract between Stanford University and the IBM Corporation. Subsequent de-

Figure 1

System Configuration for Stanford CAI *System*



velopments by IBM of the basic system have led to the IBM-1500 Instructional System, which should soon be commercially available. The basic system consists of a central process computer with accompanying disc-storage units, proctor stations, and an interphase to 16 student terminals. The central process computer acts as an intermediary between each student and his particular course material which is stored in one of the disc-storage units. A student terminal consists of a picture projector, a cathode ray tube (CRT), a light-pen, a modified

typewriter keyboard, and an audio system which can play pre-recorded messages.

The CRT is essentially a television screen on which alpha-numeric characters and a limited set of graphics (simple line drawings) can be generated under computer control. The film projector is a rear-view projection device, which permits us to display still pictures in black and white or color. Each film strip is stored in a self-threading cartridge and contains over 1,000 images each of which may be accessed very quickly under computer control. The student receives audio messages via a high-speed device capable of selecting any number of messages varying in length from a few seconds to over 15 minutes. The audio messages are stored in tape cartridges, which contain approximately two hours of messages and, like the film cartridge, may be changed very quickly. To gain the student's attention, an arrow can be placed at any point on the CRT and moved in synchronization with an audio message to emphasize given words or phrases, much like the "bouncing ball" in a singing cartoon.

The major response device used in the reading program is the light pen, which is simply a light-sensitive probe. When the light pen is placed on the cathode ray tube, coordinates of the position touched are sensed as a response and recorded by the computer. Responses may also be entered into the system through the typewriter keyboard. However, only limited use has been made of this response mode in the reading program. This is not to minimize the value of keyboard responses, but rather to admit that we have not as yet addressed ourselves to the problem of teaching first-grade children to handle a typewriter keyboard.

The CAI system controls the flow of information and the input of student responses according to the instructional logic built into the curriculum. The sequence of events is roughly as follows: The computer assembles the necessary commands for a given instructional sequence from a disc-storage unit. The commands involve directions to the terminal device to display a given sequence of symbols on the CRT, to present a particular image on the film projector, and to play a specific audio message. After the appropriate visual and auditory materials have been presented, a "ready" signal indicates to the student that a response is expected. Once a response has been entered, it is evaluated and, on the basis of this evaluation and the student's past history, the computer makes a decision as to what materials will subsequently be presented. The time-sharing nature of the system enables

us to handle 16 students simultaneously and to cycle through these evaluative steps so rapidly that from a student's viewpoint it appears that he is getting immediate attention from the computer whenever he responds.

The CAI Reading Curriculum

The flexibility offered by this computer system is of value only if the curriculum materials make sense both in terms of the logical organization of the subject matter and the psychology of the learning processes involved. Space does not permit a discussion of the rationale behind the curriculum materials that we have developed. Let me say simply that our approach to initial reading can be characterized as applied psycholinguistics. Hypotheses about the processes of reading and learning to read have been formulated on the basis of linguistic information, observations of language use, and an analysis of the function of the written code. These hypotheses have been tested in a series of pilot studies structured to simulate actual teaching situations. On the basis of these experimental findings, the hypotheses have been modified, retested, and ultimately incorporated into the curriculum as principles dictating the format and flow of the instructional sequence. Of course, this statement is somewhat of an idealization, since very little curriculum material can be said to have been the perfect end-product of rigorous empirical evaluation. We do claim, however, that the fundamental tenets of the Stanford reading program have been formulated and modified on the basis of considerable empirical evidence. There is no doubt that these will be further modified as more data accumulates.

The instructional materials are divided into eight levels each composed of about 32 lessons (1, 3, 7, 8, 11). The lessons are designed so that the average student will complete one in approximately 30 minutes, but this can vary greatly with the fast student finishing much sooner and the slow student sometimes taking two hours or more if he hits most of the remedial material. Within a lesson, the various instructional tasks can be divided into three broad areas: 1. decoding skills; 2. comprehension skills; 3. games and other motivational devices. Decoding skills involve tasks like letter and letter-string identification, word-list learning, phonic drills, and related types of activities. Comprehension involves such tasks as having the computer read to the child or having the child himself read sentences, paragraphs, or complete stories about which he is then asked a series of questions. The ques-

tions deal with the direct recall of facts, generalizations about main ideas in the story, and inferential questions which require the child to relate information presented in the story to his own experience. Finally, many different types of games are sequenced into the lessons primarily to encourage continued attention to the materials. The games are similar to those played in the classroom and are structured to evaluate the developing reading skills of the child.

The Stanford CAI Project is being conducted at the Brentwood School in the Ravenswood School District (East Palo Alto, California). There were several reasons for selecting this school. It had sufficient population to provide a sample of well over 100 first-grade students. The students were primarily from "culturally disadvantaged" homes. And the past performance of the school's principal and faculty had demonstrated a willingness to undertake educational innovations.

Computerized instruction began in November of 1966 with half of the first-grade students taking reading via CAI and the other half, which functioned as a control group, being taught reading by a teacher in the classroom. (The children in the control group took mathematics instead of reading from the CAI system.) The full analysis of the student data is a tremendous task which is still under way. However, some results have been tabulated that provide a measure of the program's success.

Some Results from the First Year

Within the lesson material, there is a central core of problems we have termed main-line problems. These are problems that each student must master in one way or another. If he passes certain screening tests, the student may bypass the main-line problems; or he may meet and solve them; he may give incorrect responses in which case he is branched to remedial material. When the first year of the project ended, the fastest students had completed 4,000 more main-line problems than the slowest students. (The cumulative response curves for the fastest, median, and slowest students are given in Figure 2.) Also of interest is the rate of progress during the course of the year. Figure 3 presents the number of problems completed per hour on a month-by-month basis again for the fastest, median, and slowest student. It is interesting to note that the rate measure was essentially constant over time for the median and slow students, but showed a steady increase for the fast student. Whether this last result is unique to our particular curriculum or will characterize CAI programs in general needs to be checked out in future research.

Figure 2

Cumulative Number of Main-line Problems for Fastest, Median, and Slowest Student

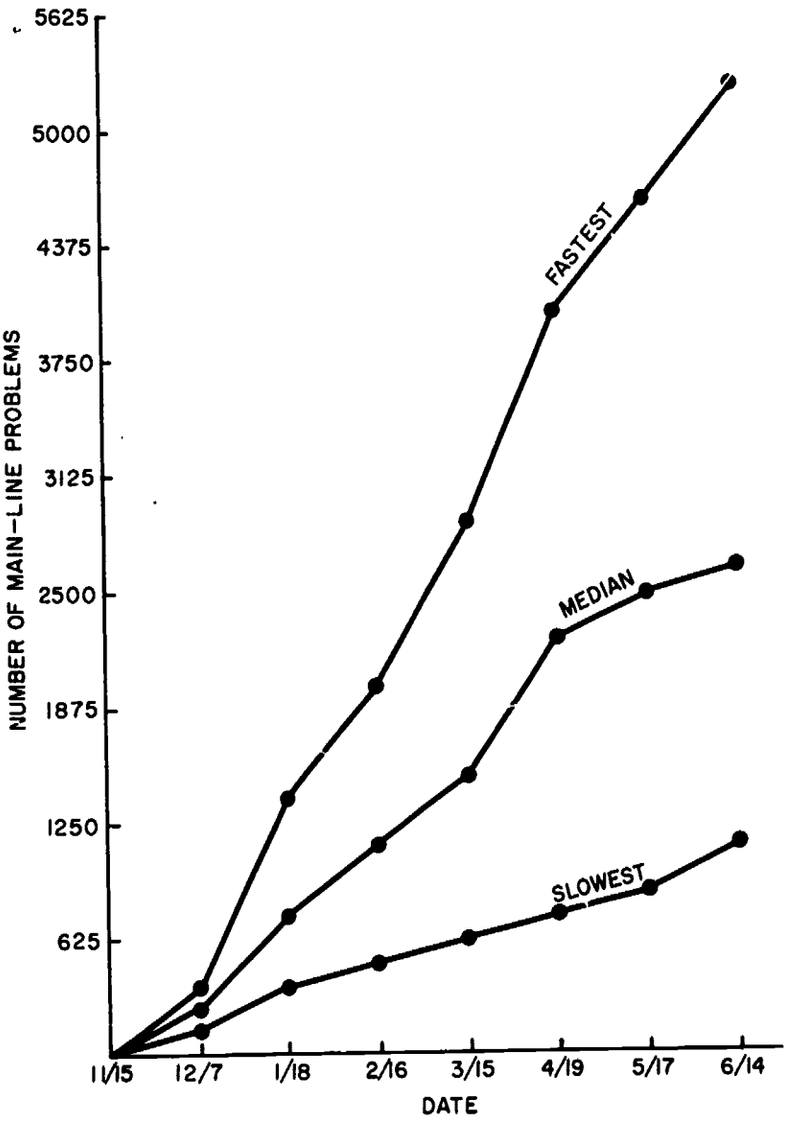
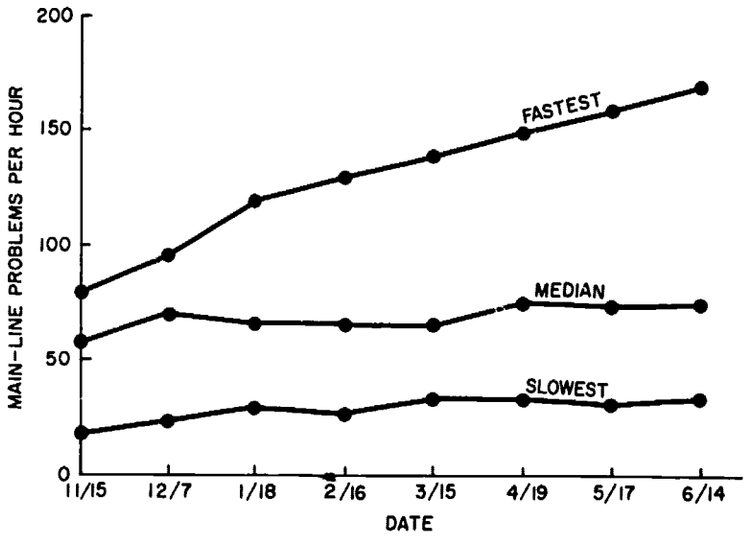


Figure 3

Cumulative Rate of Progress for Fastest, Median, and Slowest Student



From the standpoint of both the total number of problems completed during the year and the rate of progress, it appears that the CAI curriculum is responsive to individual differences. The differences noted above must not be confused with a variation in rate of response. The difference in response rate among students was very small. The average response rate was approximately four per minute and was not correlated with a student's rate of progress through the curriculum. The differences in total number of main-line problems completed can be accounted for by the amount of remedial material, the optimization routines, and the number of accelerations for the different students.

It has been a common finding that girls generally acquire reading skills more rapidly than boys. The sex differences in reading performance have been attributed, at least in part, to the social organization of the classroom and to the value and reward structures of the predominantly female primary grade teachers. It has also been argued on developmental grounds that first-grade girls are more facile in visual memorization than boys of the same age, and that this facility aids the girls in the sight-word method of vocabulary acquisition commonly

used in basal readers. If these two arguments are correct, then one would expect that placing students in a CAI environment and using a curriculum which emphasizes analytic skills as opposed to rote memorization would minimize sex differences in reading. In order to test this hypothesis, the rate-of-progress scores were statistically evaluated for sex effects. The result, which was rather surprising, showed no difference between male and female students in rate of progress through the CAI curriculum.

Sex differences, however, might be a factor in accuracy of performance. To test this notion, the final accuracy scores on four standard problem types were examined. The four problem types, which are representative of the entire curriculum, were *letter identification, word list learning, matrix construction,* and *sentence comprehension.* On these four tasks, the only difference between boys and girls that was statistically significant at the 0.05 level was for *word list learning.* These results, while by no means definitive, do lend support to the notion that when students are removed from the normal classroom environment and placed on a CAI program, boys perform as well as girls in overall rate of progress. The results also suggest that in a CAI environment the sex difference is minimized in proportion to the emphasis on analysis rather than rote memorization in the learning task. The one problem type on which the girls achieved significantly higher scores than the boys —*word list learning*—is essentially a paired-associate learning task.

As noted earlier, the first-graders in our school were divided into two groups. Half of them received reading instruction from the CAI system: the other half did not (they received mathematics instruction instead). Both groups were tested extensively with conventional instruments before the project began and again near the end of the school year. The two groups were not significantly different at the start of the year. Table 1 presents the results for some of the tests that were administered at the end of the year. These results are most encouraging. Further, it should be noted that at least some of the factors that might result in a "Hawthorne Phenomenon" are not present here; the "control" group was exposed to CAI experience in their mathematics instruction. While that may leave room for some effects in their reading, it does remove the chief objection, since these students also had reason to feel that special attention was being given to them. It is of interest to note that the average Stanford-Binet IQ score for these students (both experimental and control) is 89. While considerable variation exists, these are, by and large, not exceptional or gifted children.

Table 1

Post-tests Results for Experimental and Control Groups

Test Type	*Experimental*	*Control*	*p-value*
California Achievement Test.....	51.14	43.55	<.01
Hartley Reading Test			
Form Class.....	11.22	9.00	<.05
Vocabulary.....	19.38	17.05	<.01
Phonetic Discrimination.....	30.88	25.15	<.01
Pronunciation			
Nonsense Word.....	6.03	2.30	<.01
Word.....	9.95	5.95	<.01
Recognition			
Nonsense Word.....	18.43	15.25	<.01
Word.....	19.61	16.60	<.01

Owing to systems and hardware difficulties, our program was not in full operation last year until late in November. Initially, students were given a relatively brief period of time per day on the terminals. This period was increased to 20 minutes after the first six weeks; in the last month we allowed students to stay on the terminal 30 to 35 minutes. We wished to find out how well first-grade students would adapt to such long periods of time. They adapted quite well, and consequently this year we have, and plan to continue to use, 30-minute periods for all students throughout the year. This may seem like a long session for a first-grader, but our observations suggest that their span of attention is well over a half hour if the instructional sequence is dynamic and responsive to their inputs. Last year's students had a relatively small number of total hours on the system. This year, however, by beginning early last September and using half-hour periods throughout the year, we will be able to give each student at least 80 to 90 hours on the system.

I do not have space to discuss the sociological effects of introducing CAI into an actual school setting. There are several reports on this topic, however, and it is fair to say in summary that the reactions of

students, teachers, and parents to the program were quite favorable (1).

Nor will space permit a discussion of some of the more interesting data dealing with the evaluation of various optimization routines that are used in the reading program. In some cases, these optimization procedures are based on mathematical models of the learning processes involved and yield complex decision procedures that could only be implemented with a computer (2, 4, 6, 11). In other parts of the reading curriculum, we selected procedures that were not based on learning-theoretic considerations but were simply our best guess as to what we thought might be an optimal policy for making branching decisions among instructional materials.

Analyses of the data on optimal learning sequences have been informative and have suggested a number of experiments that need to be carried out this year. It is my hope that such analyses, combined with the potential for educational research under the highly controlled conditions offered by CAI, will lay the groundwork for a theory of instruction that is truly useful to the educator. Such a theory of instruction will have to be based on a highly structured model of the learning process and must generate optimization strategies that are compatible with the goals of education. The development of a viable theory of instruction is a major scientific undertaking, but one that cannot be ignored much longer by psychologists. Substantial progress in this direction could well be one of psychology's most important contributions to society.

REFERENCES

1. Atkinson, R. C. Instruction in initial reading under computer control: the Stanford Project. *Journal of Educational Data Processing*, 1967, *4* (on press).

2. Atkinson, R. C., Bower, G. H. and Crothers, E. J. *An introduction to mathematical learning theory*. New York: John Wiley & Sons, Inc., 1965.

3. Atkinson, R. C. and Hansen, D. N. Computer-assisted instruction in initial reading: the Stanford Project. *Reading Research Quarterly*, 1966, *2*, 5-25.

4. Atkinson, R. C. and Shiffrin, R. M. Human memory: a proposed system and its control processes. In K. W. Spence and J. T. Spence (Eds.),

The psychology of learning and motivation: Advances in research and theory, Vol. 2. New York: Academic Press, 1968 (on press).

5. Fishman, Elizabeth J., Keller, L. and Atkinson, R. C. Massed vs. distributed practice in computerized spelling drills. Technical Report 117, Institute for Mathematical Studies in the Social Sciences, Stanford University, 1967.
6. Groen, G. J. and Atkinson, R. C. Models for optimizing the learning process, *Psychological Bulletin*, 1966, *66*, 309-320.
7. Hansen, D. N. and Rodgers, T. S. An exploration of psycholinguistic units in initial reading. Technical Report 74, Institute for Mathematical Studies in the Social Sciences, Stanford University, 1965.
8. Rodgers, T. S. Linguistic considerations in the design of the Stanford computer-based curriculum in initial reading. Technical Report 111, Institute for Mathematical Studies in the Social Sciences, Stanford University, 1967.
9. Suppes, P. The uses of computers in education. *Scientific American*, 1966, *215*, 206-221.
10. Suppes, P., Jerman, M. and Groen, G. J. Arithmetic drills and review on a computer-based teletype. *Arithmetic Teacher*, April, 1966, 303-308.
11. Wilson, H. A. and Atkinson, R. C. Computer-based instruction in initial reading: A progress report on the Stanford Project. Technical Report 119, Institute for Mathematical Studies in the Social Sciences, Stanford University, 1967. (To be published in *Basic studies in reading*, edited by H. Levin and Joanna Williams. New York: Harper & Row.)

Academic Games and Learning

JAMES S. COLEMAN
Johns Hopkins University

My aim in this paper is to give some insight into what academic simulation games are, what their goals are, and how they accomplish these goals. I want to describe how simulation games differ from other ways of teaching and learning—both in the way children learn from them, and in the kinds of things they learn. As will become evident, these differences are sharp ones indeed, and I will count it sufficient achievement to do no more than communicate them.

The Relation between Games and Learning

A "simulation game" combines the properties of games in general with the properties of simulations in general. The essential properties of a game for present purposes are these: 1. Its basic elements are players or actors, each striving to achieve his goal; 2. it is limited to a small, fixed set of players; 3. its rules limit the range and define the nature of legitimate actions of the players; 4. again, through the rules, it establishes the basic order, sequence, and structure within which the actions take place; 5. it is delimited in time as well as extensivity, with an end defined by the rules; and 6. its rules constitute a temporary suspension of some of the ordinary activities of life and rules of behavior by substituting for them these special time-and-space delimited ones.

In short, a game is a way of partitioning off a portion of action from the complex stream-of-life activities. It partitions off a set of players, a set of allowable actions, a segment of time, and establishes a framework within which the action takes place. It establishes what one might describe as a minute system of activities, and if the game contains more than a single player (as most games do), the game can even be de-

scribed as a minute social system.

It is undoubtedly for this reason that games are such an important part of the socialization of young children. For the playing of a game allows a child to practice, in this limited framework, action that is interdependent with the actions of others, carried out within a set of rules, and in pursuit of a goal. As Piaget's observations of children playing the game of marbles show, children do not immediately learn the idea of playing a game, and only slowly gain a sense of the nature of its rules. Piaget suggests that the learning of the nature of rules in a game is, in fact, the learning of the nature of a moral order.

Thus, games may be regarded as a special invention in which children or adults practice with the components of life itself, a kind of play within the larger play of life. Because they are constructed of these components of life, games as means by which children learn deserve more serious attention than they have received. It is true, to be sure, that games are used by teachers in early grades of school, both as general instruments of socialization and as vehicles for teaching certain content. But they are generally regarded as auxiliary aids to the essential task of "teaching," and after the early elementary grades, are forsaken in favor of more serious approaches to teaching.

These more serious approaches to teaching are based on a very different conception of how children can, or perhaps should, learn. This conception is one based on the idea of *transmission* of knowledge (or skills, or ideas) from a teacher to a student followed, in some cases, by practice of the student in the repetition or use of this knowledge, skill, or idea. It has many variants, including transmission through a variety of media, such as audiovisual aids, books, educational television, and others. But the basic model is the same: a conception of the child as a receptor of knowledge, skills, or ideas transmitted from others.

It is only by contrasting these two models of a learning context that one begins to see the rather peculiar characteristics of this second, or school, model of learning. The school model has none of the remarkable lifelike properties that a game has, but appears to be a simplistic use of the fact that information is transmitted by communication, and that repetition aids learning.

The comparison of the school model of learning with experimental and theoretical work by psychologists makes this model appear even more puzzling. In this work, the two essential properties of the learning context are action in an environment and reward; the learner is always *learning to act by acting*. Furthermore, it is important to note that the

learning is *incidental* to his goal; the goal is not learning itself. The student is motivated to receive the reward; he learns a given action only because it is this action that gains him the reward.

The learning that occurs is a kind of "learning to be motivated" in a given direction, learning to generalize his affect from one stimulus or environmental context to another. Once he has "learned" or "become motivated," then he may pick up more and more information about the new environmental context that enables the action to take place more efficiently, and certainly this might be called learning as well. But the essential step is the development of affect toward the new environmental context, or to put it another way, learning to be motivated—*i.e.*, to act, in a new direction.

This language contrasts sharply with that used to describe the classroom. When children fail to "learn their lessons," it is often said that they are "not motivated to learn," and consequently cannot be taught. The task is regarded as one of "teaching" children *after* they have already been motivated. Thus, while psychologists consider the most essential step as *learning to be motivated to act in a given direction, to achieve a given goal,* the school is seen to operate under the assumption that a child *is already motivated* to learn mathematics or history or English literature. Consequently, all that is necessary for the teacher is to provide that information that facilitates his movement toward these goals. Obviously the child will assimilate it because—the implicit argument goes—the information he has been provided with does indeed facilitate reaching that goal.

Viewed in this light, it becomes much more evident why variations in school seem to have so little effect on what a child learns compared with variations in his family background. For if the essential learning task is that of transfer of affect—or learning to be motivated (that is, to act) toward the object—then this has been carried out in the home prior to, and concurrent with. the school, but not within the school.

There is, of course, one way that schools give many children a motivation that partially coincides with the goal of learning mathematics or history or English literature. This is by establishing grades and a diploma. The child in many homes "learns to be motivated" toward the goal of good grades by his parents' ability to transfer his affect to this goal. But as every teacher knows, this goal is only partially coincidental with that of knowing mathematics, or history, or English literature. Moreover, not every child is given this goal by his family.

In short, it appears that the usual conception of the school's task

leaves out the most crucial step in learning: the necessary and almost sufficient condition of developing strong affect toward goals that require the content the school teaches—that is, learning to act toward these goals.

It is within this framework that I want to examine the characteristics of games as learning tools. For I suggest that playing a game with a given content has precisely the effect of "learning to be motivated" toward assimilating that content. The game provides the goal for which the content is relevant, and the very nature of games insures that the player will be motivated toward that goal. I suggest that the game fulfills precisely the step that is missing in the usual conception of a school's task—the learning that leads a child to actively assimilate the information transmitted to him in school. It is true that some good teachers, particularly those in elementary grades with enough time to give attention to individual children, recognize the need for this step, and by improvising, attempt to carry it out. But the essential formal task of the teacher is seen according to the simplistic model described earlier: "teaching" by transmitting information. If my points above are correct, then providing such information is only the second step; the first is to bring about the true learning—the learning to act or be motivated toward a goal which the information facilitates.

The kind of learning that can go on in a game, then, is complementary to, and logically prior to, the kind of learning that occurs in the standard information-transmission model of school learning. Learning in a game is the development of affect toward a new goal; and the transmission of knowledge that occurs in an ordinary classroom is a way of facilitating action toward that goal. From this perspective, and I suggest from the learner's perspective as well, he is not carrying out actions in order to assimilate the material presented to him. It is quite the reverse; he is assimilating the material in order to be able to efficiently carry out actions toward his goal. The goal may be a goal in a game toward which this content is relevant or, less likely, it may be a goal in real life toward which the content is relevant. Most frequently, of course, it is the goal of getting good grades, toward which the content has no logical relevance, but which the school has artificially connected to the content.

This perspective, if correct, implies a number of points about the use of games for learning in schools. First, it implies that the appropriate games for learning are those in which winning, or attainment of the goal, is in fact facilitated by the knowledge that the school is attempting

to "teach." Games with goals unrelated to such content will not in themselves make the child "motivated to learn."

Second, the sequence of game-learning and information-transmission should obviously be such that the game is first prior to, and then interspersed with, the information-transmission. For the goal must be learned in order for the information to be relevant, and the goal must persist so that the information continues to be relevant.

Third, this perspective implies something about the relative amounts of attention that the school should give to the two kinds of learning—that is, that the goal-learning should receive the greater amount of time and effort. There are many examples of students avidly seeking out information in order to do better in achieving the goals of a game; but I know of no examples in which students, given new information, go out and seek goals which this information could facilitate. Incidentally, this appears to be the principal reason that graduate school training is a remarkably affective resocialization process. The activities of the faculty are designed more to induce motivation toward new goals than they are to transmit information. In a good graduate department, the students get the information on their own, once they have learned motivation toward the new goals (a task which is facilitated by the fact that their present teachers are their future colleagues and judges throughout a career).

Fourth, this perspective about goal-learning and information-transmission implies that the most direct and powerful impact of games in schools will be upon children described as "unmotivated." For these children have never learned a goal to which school is relevant. The effect on children who are already "highly motivated" should be more subtle, less directly upon the overall amount of achievement, more on the style of activity and the profile of achievement. For example, games should lead them to a more uneven profile of achievement as they learn one set of goals more fully than another, and thus seek out information on the first more avidly than the second. This subtle effect may, however, have long-run consequences because the goals of a game and the content of school have a direct and logical coherence. The goal of good grades, which motivates most "highly motivated" children, has no necessary relevance to the content of school. Thus, when these children graduate, and the goal of good grades no longer obtains, there is no related goal to support that information and motivate its expansion. The result is not merely that the former students quickly forget information that is irrelevant to their current goals, but something consider-

ably more important: They have never learned a goal at all. One result, of course, is the curious disorientation that occurs for many adolescents at the end of high school, especially for those upper-middle class adolescents who have never learned economic goals from economic necessity.

Games and the Learning of Structure

A second way in which learning through games differs from the school model of learning derives from the properties of games described earlier. A player's role in a game consists of a structure of interrelated actions toward a goal. Learning of this structure of actions, and their relation to the larger structure of actions of all the players, constitutes learning both the whole and the relation between the parts. This structure of action, once learned, becomes a structure to which relevant information is assimilated. Thus the information, when it is assimilated, is not merely "learned"; it is fitted into the structure of action in such a way that it facilitates achieving a goal. Thus the game provides the structure which Bruner argues is so important to retention and usability of information. The structure learned in this way is even more deeply embedded than one that is learned only cognitively.

It is very likely that one reason education in schools proceeds as well as it does in subjects like mathematics is that the school model I've described is used less often in that subject area. In solving arithmetic, algebraic, and other mathematical problems, the child is himself engaged in a small game with a well-defined goal. He learns mathematical operations through the action of employing them toward that goal. This sometimes fails, of course, because the "game" that is set for him is sometimes too hard, and he never reaches the goal but merely experiences failure; if and when he does reach it, however, the means by which he is learning has many similarities to learning through games.

Games and the Human Sciences

One of the reasons social studies is so poorly taught in high schools is that the schools have few, if any, means for providing the appropriate structure within which it should be learned—a structure of human action. What a simulation game in the area of social studies does is provide such a structure of action, one within which the information the student learns can be located and fixed in his memory. It may well be, in fact, that simulation games are more appropriate to social studies

than to other subjects for just this reason. For social studies involves the actions of human actors; and the playing of a game embeds in one's experience that particular structure of action.

Simulation games can also be devised for the physical sciences; but in those disciplines, the relevant actions are those of the physical environment, which can be as well observed in a laboratory experiment as in an interpersonal game. However, the game does lend some things that an experiment does not, such as the added motivation that occurs when a number of persons are striving toward interdependent goals. This motivation is not trivial, as evidenced by the success of some mathematical games, such as Wff-n-Proof, the logic game, and Equations, a game involving the creative use of arithmetic operations. But apart from these values of games in other intellectual domains, the isomorphism between the very structure of games and the human sciences is striking indeed and suggests their special values for these areas of learning.

What Is a Simulation Game Like?

One of the games developed and used by the Johns Hopkins Academic Games Project, under a grant from the Carnegie Corporation, should give some idea of what such games are and do. It may be described as a legislative game and is played as follows: A group of 6 to 13 players constitutes a legislature, and the game is a session of the legislature in which eight issues are introduced and voted upon. Each player is dealt cards each of which shows the preferences of his constituents on a certain issue.

Each player has as his goal the simple task of getting reelected. But to accomplish this, he must get as many issues passed (or defeated) as he needs to satisfy the majority of his constituents. The votes for and against him in reelection after the bills are passed are determined by the numbers of his satisfied and dissatisfied constituents and the outcome of each issue as shown on the faces of the cards.

This structure of the game induces, as one might expect, a variety of negotiations, vote exchanges, and bargains of various sorts by each player in order to gain control of the outcome on those issues important to his constituents. Thus, the principal kind of action that the player engages in is one of the kinds of actions that real legislators engage in. The player comes to see the connection between the legislator's constituency and the legislator's actions and the connection between the legislator's goals of reelection and the kind of behavior he carries out.

A number of points can be made about this game. First, the players learn information relevant to the game: They do not go out and learn information about the content of issues, they go out and learn information about the functioning of a legislative body. (Players have learned, for example, the fine points of Robert's Rules of Order to facilitate their gaining reelection.) Although their *interest* in the content of the issues is stimulated, and there is evidence that their attitudes change somewhat on these issues, they do not seek out information on this content. The game that would induce them to seek out such information would be a different game, one in which information about the actual content of the issues facilitated achieving their goals.

A second point is that the structure of this game very selectively abstracts a single process of negotiation and bargaining that occurs in legislatures—an important process, but not the only one. It does so because the learning of this process can occur unencumbered by the additional processes of which real legislatures are composed. The additional processes are learned in a stepwise fashion as each player encounters different levels of the game. At each higher level, an additional process is introduced such as committee structure; introducing the importance of the legislator's own values concerning an issue in addition to his goal of reelection; introducing special powers for the floor leader or chairman; and additional complexities. Thus, the complex structure of a legislature is learned by first analytically separating the various processes in it, and then reconstructing the functioning legislature in a stepwise fashion.

A third point is that the goal of the player in the game, and the constraints on his behavior, are made as nearly like those of the real actor in the situation as possible, subject to the conditions described under the second point above. Thus, the simulated structure of action is designed to mirror, so far as possible, the motives and interests of a real person in such a situation. The structure of action which is learned and which constitutes the framework into which information is fitted is like that in reality. The player, as a consequence, has a natural screening device for information, and a natural basis for choosing what information to seek out. The game is a good simulation of reality because he seeks only the information he would need for acting in this kind of situation. He does not learn the information the teacher says is important or that which he thinks will give him a good grade, but the information he will need for action.

A fourth point is that the general principles exhibited in playing the

game are not recognized in verbalizable form by all players. Some players quickly infer the general principle of interdependence between legislators and constituents which makes legislatures function; others do not learn this until discussions following the game. Virtually all understand this after such discussions. But the phenomenon which has been observed in other contexts—that some persons translate their experience into general principles which they can verbalize, while others do not—applies here as well. As a consequence, for many children, a strong, second learning experience occurs in discussions after the game. This point illustrates the more general point made earlier: That play in a game is not a self-contained learning method but one that is complementary to the verbal discussions and information-transmission of which most school activity is now composed.

A final point about this game and others is the wide range of skill and background they encompass. The game has been played by seventh graders and by graduate students, in identical form. It has been played by students in a ghetto school and has provided the basis for at least two faculty members' theoretical papers on the topic of legislative decisions. This broad span is not merely characteristic of this game but of simulation games in general. The practical implications of this are enormous, of course. Today, the span of ability that can be encompassed by current teaching methods is so narrow that schools must resort to tracking and grouping. The evidence from use of games in the classroom indicates that a much broader span of ability can be usefully encompassed by simulation games.

Conclusion

I want to be very clear about what I am suggesting in this paper concerning simulation games and learning. I am not describing games as a "new teaching device"; I am rather suggesting that the use of games in learning introduces fundamental changes in the nature of the task the school is carrying out; that the use of games constitutes a fundamental change in the process by which learning takes place; and that the intrinsic character of games means that simulation games are especially appropriate for embedding into experience and cognition the structure of social action on which the human sciences are based. Certain of the details of these arguments may be incorrect; but they constitute a strong challenge to the current teaching activities of schools, especially in the area ordinarily termed social studies.

Luncheon Address

Testing and Public Policy

WILLIAM GORHAM
U.S. Department of Health, Education, and Welfare

Shortly after I began my current assignment, I discovered that the Department of Health, Education, and Welfare (HEW) offers a distinct advantage to the newcomer, an advantage which derives from the diversity of its interests and constituencies. In describing the work of my office, I have found it both possible and necessary to draw examples from fields outside my listeners' expertise—health examples for educators, income-maintenance examples for health professionals, and so on. If I talk today a little less about testing than you might wish, you could attribute it to force of habit. But I beg you not to do so. Instead, regard it as a genuine desire on my part to share with you two warm conclusions I have reached while trying to improve the basis for allocating monies among the public programs of HEW.

The first conclusion is that right now we (and by "we" I do mean all of us) do not have the means to make informed choices among alternative ways of spending money on education. We lack information. Specifically, we lack information about the relationship between student performance and the resources that go into education. I will not belabor this conclusion by cataloguing the things that we don't know. Take my word for it, we don't know much for sure, and on some crucial questions (such as effective compensatory education) it's fair to say that we don't have enough solid information to plan ourselves out of a lunch bag.

The second conclusion is that our best hope for improving the performance of the educational system lies in improving the incentive and the capacity of the system itself to learn.

I can hardly launch into this subject without defining what I mean

by "improving education." At this conference a year ago, Henry Dyer presented a paper on educational goals. It was a very good paper. He emphasized the importance of paying attention to *all* outcomes of schooling. Some of the outcomes that we now ignore (perhaps because they do not appear as curriculum objectives are extremely important. The example he used was the fact that a very large fraction of high school kids cheat. The alarming point was that a very large fraction of the cheaters think cheating is perfectly all right. A society which ignores this sort of outcome of its educational process is halfway down the drain already.

Another outcome which should demand equal attention is self-esteem—the way kids feel about themselves and their ability to influence their future. If our educational system teaches young people how to read and write, teaches them history and geography, but somehow undermines their sense of their own worth and capacity, we may be paying too high a price for the three R's.

Measuring these intangible but crucial outcomes as by-products of the educational system is, in my judgment, among the most important challenges to the testing community. I should like to discuss briefly a number of questions that are related to these two conclusions: Who makes the decisions about how to improve education? What kind of information do they need? Why don't they have this information now? What can be done about it? And by whom?

Who makes the decisions that lead to better education? These decisions are made at all levels. They are made by teachers every day all over the country. Principals also make these decisions every day; so do superintendents. Less frequent, but far-reaching, decisions that may change the form as well as the substance of education are made by school boards, state agencies, legislatures, federal agencies, and the Congress.

Let us start with the teacher: How does a good teacher improve education in his classroom? Mostly he does it by experimenting and observing the results of his experiments. He gets ideas from everywhere—from his own training and background, from reading, from observing other teachers, from demonstration projects. He operates under constraints, often severe ones. Mostly these constraints are the various resources at his disposal: his own time, the curriculum, the equipment. There are many decisions he is not permitted to make, but within the boundaries of his influence, he is constantly trying different ways of teaching, then evaluating his results. He tries new material as well as

new ways of presenting old material.

He isn't seeking a "best" method of teaching. Individuals differ, and classes differ in their dynamics, even when they contain apparently similar mixtures of age, sex, and abilities. Moreover, he knows that newness itself is often necessary. If he presents the same material again and again in the same way, he will end by boring himself and his students. The Hawthorne effect is his ally. The good teacher discovers by experience that one book is generally more exciting than another; that one method of presentation works better, either in general or with particular types of children—with boys, with girls, with slow students, with fast students, and so forth.

The teacher has an enviable feedback system with a wide range of measures of success. Test scores are available, but a good teacher never relies on test scores alone. He has available to him other measures of success: student attitudes, enthusiasm, the degree of shine in the eye, the willingness to work independently and creatively. The good teacher continuously measures success in these ways, but he does not use the same scale for each child. Intuitively, he tries to correct—but not overcorrect—for variables he cannot control: ability, family background, previous achievements, physical handicaps, and so forth.

I have described the familiar image of how the good teacher makes decisions for improving education in his classroom because I think it is highly relevant to "system" learning and decisions made at higher echelons. Principals, superintendents, and school boards make different kinds of decisions. They decide how to allocate a budget among different levels of education and different kinds of resources; when to seek funds for a new building; when to raise teachers' salaries in order to attract more teachers or more qualified teachers. They decide what is meant by higher qualifications and what mix of teachers and nonprofessional personnel is desirable. They make decisions about curriculum and test books and equipment that affect a whole school or a whole school system. The extent of their influence for making education better or worse is very large. Their decisions affect many more children than the decisions of the individual teacher because they set most of the constraints within which teachers must perform.

How does a good administrator (principal, superintendent, school board member) make decisions that will improve education? The ideal school administrator acts very much like the good teacher. He is continuously evaluating variation in the grist and mill of education. He tries different things in different schools; hiring different kinds of teach-

ers, trying different curricula, different methods, different class sizes, different schedules, *and* he "observes" what happens. The best administrator has some variations going almost all the time. Sometimes he tries major departures. If he isn't occasionally in hot water with his community, he isn't doing his job.

How would this ideal administrator measure success? He is one or two steps removed from the tell-all faces of the children and therefore he cannot get the immediate feedback that makes a teacher's information system so enviable. The administrator does, however, have access to some measures of success: achievement test scores, dropout rates, attendance rates, continuation rates. There is little direct evidence on attitudes and motivation, but he could have a variety of attitude measures developed. He cannot evaluate these qualities intuitively as a teacher can, nor can he so easily correct for noncontrollable variables. He needs a more systematic and formal feedback mechanism; he needs information about, and analysis of, the relations between the various measures of success and the specific characteristics of the education being given to specific children. Such a system doesn't come naturally no matter how gifted the administrator. He has to build some parts of it from scratch.

Actually very few, if any, school and system administrators act the way I have described. Uniformity rather than intentional variation is the rule, and systematic analysis of anything important in education is almost nonexistent. In general, variation isn't planned; it creeps into a school system. New schools have libraries and lunch rooms, old ones don't; schools in better neighborhoods have more experienced and better trained teachers because these teachers get first choice of jobs. Because this unplanned variation could be politically embarrassing, most school systems would rather not even *know* about it. They do not keep the books in such a way that the variations in resources by school, for example, can be easily discovered.

Even where substantial variation occurs—in the mix of resources, the curriculum, or the method of teaching (by accident, by default, or even by design)—the effects of this variation on the performance of the children is rarely analyzed. It is nothing short of incredible that the test scores and other measures of performance of children, so laboriously and expensively collected by most school systems, are almost never used to give clues to the relative effectiveness of different types or conditions of education. Test scores are used by teachers to grade students and gauge their progress. Broad averages may be reported for

the system as a whole to give a general picture of change from year to year. But no school system that I know of actually analyzes how changes in test scores are related to changes in the method of education or to resources devoted to particular children. If they were analyzed, they might provide us with guidance for improving the educational system.

Planned variations and analysis of results don't happen for a number of reasons:

- Because it's "unfair." One must certainly sympathize with the school administrator who is attacked from all sides—by teachers as well as by parents—if he seems to be treating any group differently from any other group. He might very much want to experiment with radical changes in class size, moving to classes of 10 in School A and classes of 40 in School B, and evaluating the results over a period of years. But you can imagine the "unfair" cry he would receive from parents and teachers in School B. He is unlikely to take the risk. Even if the total resources devoted to each child were equal in the two schools, parents in both schools would complain that the method being tried in the other school was better and why weren't their children getting it. Until the necessity for variation and experimentation is well understood by parents and teachers as well as administrators, only the rare administrator will take the risk of offending.
- Because it's risky. It is of the essence of experimentation that some experiments don't work. If one is experimenting with physical substances, the cost of failure is time and money. If one is experimenting with children, the cost of failure may be very great. A group of children exposed to a new method of teaching reading may not learn to read. They may feel themselves to be failures because other children exposed to some other method are already reading.
- For other reasons. Most school administrators do not have the incentive, the resources, or the know-how to build into their systems a capacity for systematic institutional learning. Their training and experience typically do not motivate or equip them to think in this way. And they really *must* be motivated because they have barriers to overcome—an unsympathetic or complacent community, conflicting demands for resources (for tangible, no-nonsense things like gymnasiums or school band uniforms), staffs or colleagues as difficult to activate as the most conservative elements of the community, and so on.

In short, the ideal administrator, as I have described him, will not be without scars. But let us leave him for now and discuss briefly the more abstract role of the Federal Government.

The problems confronting the federal policy maker in education are similar to those faced by educators at lower levels: how to use limited aid to state or local governments; aid for construction or equipment or libraries; teacher training and research. The federal resources could be concentrated on preschool or the early years, as in HEAD START and Follow-Through, or they could be concentrated at other ages, or spread over all ages.

To make good decisions (*i.e.*, to choose a mix of policies that will contribute more to the improvement of American education than other mixes of policies), federal policy makers need some indication of the relationship between the addition of resources, the use of those resources, and the performance of students. Will buildings help more than the raising of teachers' salaries? Will poor children be benefited more by preschool than by teen-age programs?

For practical purposes, the federal policy makers are as far from the shining eyes of children as from the stars. They get no sensory, and little other, feedback when they try out policies.

Where does the federal policy maker turn for some clues to the relationships between uses of resources and measures of success in student performance?

1. *Statistical surveys.* Project TALENT, the Equal Opportunity Survey (the Coleman study), and others have analyzed the statistical relationships between inputs (teachers, buildings, materials, and so forth) into the educational system, the characteristics of students, and the performance of these students on tests. These surveys have given important information about the actual performance of different kinds of children and the resources devoted to their education. They have also shown, however, that most of the variation in children's performance is attributable to differences in family background. When family factors are held constant, a clear picture of the relationship between school variables and student performance does not emerge. The Coleman study does suggest, however, that verbal ability of teachers is an important input.

   The results of these surveys have been interesting, often highly significant, but *not* very much help to the federal policy maker anxious to make the right decision about the use of federal resources.

We have a long way to go to learn all that we can from surveys but, in my opinion, they will tell us little about the relative effectiveness of alternative uses of resources because of the "noise" of uncontrolled variables.

2. *Project evaluation.* In the past year, a major effort to evaluate individual projects under HEAD START, Title I, and other federally financed programs has been launched with a view to seeing what kinds of methods work well under what conditions. This is uphill and unpromising evaluation since the programs in question were not designed for "learning." Thus far, it has been virtually impossible to isolate the causes of change—or stability—in achievement test scores. If state and local educational authorities were consciously evaluating educational programs, it might not be so important for the Federal Government to do this. But until state and local educational systems begin to experiment more widely, I believe the Federal Government must take the lead. It must use federal resources actively to encourage variation and evaluation of this variation. Programs like HEAD START and Title I provide great opportunities, not just to serve the children they reach but to serve all children.

The Federal Government can stimulate experimentation, can plan its programs to promote systematic variation and evaluation, and can subsidize and encourage research and disseminate the results. But let's not kid ourselves about the possibility of a revolution from Washington. While the creative use of federal funds can provide the incentive needed to spur the system itself to learn—by developing new techniques, by stimulating the evaluation of old ones, and by disseminating the products of these efforts—real progress toward improving education will come only when a substantial number of teachers and educational administrators at all levels see themselves as involved in a continuous learning process. If the Federal Government succeeds in bringing this about, it will be able to retreat to the satisfying role of the rich uncle.

Session III

Theme:
Measurement
Systems

Sample-free Test Calibration and Person Measurement

BENJAMIN D. WRIGHT
University of Chicago

My topic is a problem in measurement. It is an old problem in educational testing. Alfred Binet worried about it 60 years ago. Louis Thurstone worried about it 40 years ago. The problem is still unsolved. To some it may seem a small point. But when you consider it carefully, I think you will find that this small point is a matter of life and death to the *science* of mental measurement. The truth is that the so-called measurements we now make in educational testing are no damn good!

Ever since I was old enough to argue with my pals over who had the best IQ (I say "best" because some thought 100 was perfect and 60 was passing), I have been puzzled by mental measurement. We were mixed up about the scale. IQ units were unlike any of those measures of height, weight, and wealth with which we were learning to build a science of life. Even that noble achievement, 100 percent, was ambiguous. One hundred might signify the welcome news that we were smart. Or it might mean the test was easy. Sometimes we prayed for easier tests to make us smarter.

Later I learned one way a test score could more or less be used. If I were willing to accept as a whole the set of items making up a standardized test, I could get a relative measure of ability. If my performance put me at the eightieth percentile among college men, I would know where I stood. Or would I? The same score would also put me at the eighty-fifth percentile among college women, at the ninetieth percentile among high school seniors, and above the ninety-ninth percentile among high school juniors. My ability depended not only on *which* items I took but on *who* I was and the company I kept!

The truth is that a scientific study of changes in ability—of mental

development—is far beyond our feeble capacities to make measurements. How can we possibly obtain quantitative answers to questions like: How much does reading comprehension increase in the first three years of school? What proportion of ability is native and what learned? What proportion of mature ability is achieved by each year of childhood?

I hope I am reminding you of some problems which afflict present practice in mental measurement. The scales on which ability is measured are uncomfortably slippery. They have no zero point and no regular unit. Their meaning and estimated quality depend upon the specific set of items actually standardized and the particular ability distribution of the children who happened to appear in the standardizing sample.

If all of a specified set of items have been tried by a child you wish to measure, then you can obtain his percentile position among whatever groups of children were used to standardize the test. But how do you interpret this measure beyond the confines of that set of items and those groups of children? Change the children and you have a new yardstick. Change the items and you have a new yardstick again. Each collection of items measures an ability of its own. Each measure depends for its meaning on its own family of test takers. How can we make objective mental measurements and build a science of mental development when we work with rubber yardsticks?

Objectivity in Mental Measurement

The growth of science depends on the development of *objective* methods for transforming observation into measurement. The physical sciences are a good example. Their basis is the development of methods for measuring which are specific to the measurement intended and independent of variation in the other characteristics of the objects measured or the measuring instruments used. When we want a physical measurement, we seldom worry about the individual identity of the measuring instrument. We never concern ourselves with what objects other than the one we want to measure might sometime be, or once have been, measured with the same instrument. It is sufficient to know that the instrument is a member in good standing of the class of instruments appropriate for the job.

When a man says he is at the ninetieth percentile in math ability, we need to know in what group and on what test before we can make any

sense of his statement. But when he says he is five feet eleven inches tall, do we ask to see his yardstick? We know yardsticks differ in color, temperature, compositions, weight—even size. Yet we assume they share a scale of length in a manner sufficiently independent of these secondary characteristics to give a measurement of five feet eleven inches objective meaning. We expect that another man of the same height will measure about the same five feet eleven even on a different yardstick. I may be at a different ability percentile in every group I compare myself with. But I am the same 175 pounds in all of them.

Let us call measurement that possesses this property "objective" (2, 4, 5). Two conditions are necessary to achieve it. First, the calibration of measuring instruments must be independent of those objects that happen to be used for calibration. Second, the measurement of objects must be independent of the instrument that happens to be used for measuring.* In practice, these conditions can only be approximated. But their approximation is what makes measurement objective.

Object-free instrument calibration and instrument-free object measurement are the conditions which make it possible to generalize measurement beyond the particular instrument used, to compare objects measured on similar but not identical instruments, and to combine or partition instruments to suit new measurement requirements.**

The guiding star toward which models for mental measurement should aim is this kind of objectivity. Otherwise how can we ever achieve a quantitative grasp of mental abilities or ever construct a *science* of mental development? The calibration of test-item easiness must be independent of the particular persons used for the calibration. The measurement of person ability must be independent of the particular test items used for measuring.

When we compare one item with another in order to calibrate a test, it should not matter whose responses to these items we use for the comparison. Our method for test calibration should give us the same

*There is a third condition which follows from the first two. The evaluation of how well a given set of observations can be transformed into objective measurements must be independent of the objects and instruments that are used to produce the observations. It must also be reasonable to hypothesize that objects and instruments have stable characteristics which do not interact with each other.

**Were it useful to glue three 12-inch rulers together to make a 36-inch yardstick or to saw a 36-inch yardstick in three to make some 12-inch rulers. we would retain our confidence in the objective meaning of length measurements made with the resulting new instruments.

results regardless of whom we try the test on. This is the only way we will ever be able to construct tests which have uniform meaning regardless of whom we choose to measure with them.

When we expose persons to a selection of test items in order to measure their ability, it should not matter which selection of items we use or which items they complete. We should be able to compare persons, to arrive at statistically equivalent measurements of ability, whatever selection of items happens to have been used—even when they have been measured with entirely different tests.

An Individualistic Approach to Item Analysis

Exhortations about objectivity and sarcasm at the expense of present practices are well and good. So what? Can anything be done about it? Is there a better way?

In the old way of doing things, we calibrate a test item by observing how many persons in a standard sample succeed on that item. Item easiness is defined by the proportion of correct responses in the sample. Item quality is estimated from the correlation between item response and test score. Person ability is defined by percentile standing in the sample. This approach leans heavily on assumptions concerning the appropriateness of the standardizing sample of persons.

A different approach is possible, one in which no assumptions need be made about the persons used. This new approach assumes instead a very simple model for what happens when any person encounters any item. The model says simply that the outcome of the encounter is governed by the product of the ability of the person and the easiness of the item. Nothing more. The more able the person, the better his chances for success with any item. The easier the item, the more likely any person is to solve it. It is as simple as that.

But this simple model has a surprising consequence for item analysis. When measurement is governed by this model, it is possible to take into account whatever abilities the persons in the calibration sample happen to have and to free the estimation of item easiness from the particulars of these abilities. The scores persons obtain on the test can be used to remove the influence of their abilities from the item analysis. The result is a person-free test calibration.

I learned this kind of item analysis from the Danish mathematician Georg Rasch. But comparable work has been done here by Frederic Lord and Allan Birnbaum. Some of the ideas have been in print for

years. What is surprising is that this powerful method is not used in practice.

Why not? Perhaps too few recognize the importance of objectivity in mental measurement. Perhaps, too, many despair that it can ever be achieved, or fear it will be too difficult to do. What we need is some evidence that objective measurements of mental ability can really be made.

The crucial questions are: Can test calibration really be independent of the ability characteristics of the persons used to make the calibration? Can person measurement, the estimation of a person's ability from a score on some selection of test items, really be independent of those items used for the measurement?

The data in this paper illustrate that both of these ideals can be lived up to in practice. These data happen to come from the responses of 976 beginning law students to 48 reading comprehension items on the Law School Admission Test. But they are only one illustration. The method has also worked with other mental tests (2).

Person-free Test Calibration

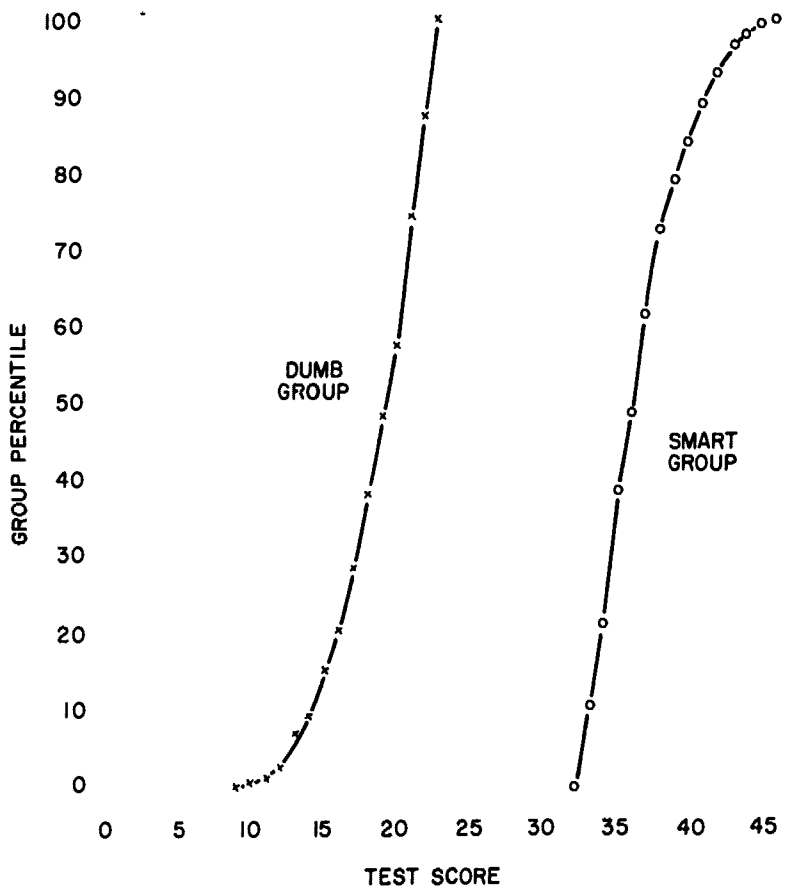
In order to examine the dependence of test calibration on the abilities of these law students, let us construct the worst possible situation. Into a Dumb Group we will put the 325 students who did worst on the test. The best of them got a score of 23. Into a Smart Group we will put the 303 students who did best. The worst of them got a score of 33. Thus, we have two groups dramatically different in their ability to succeed on this test of reading comprehension. There are 10 points difference between the smartest of the Dumb Group and the dumbest of the Smart Group.

Now for the acid test. How would a test calibration based on the Dumb Group compare with one based on the Smart Group? To remind us of how things look using the old way of doing things, I made up these calibrations in terms of sample percentiles. Each curve in Figure 1 represents a person-bound test calibration. The curve on the left is the test calibration produced by the Dumb Group. The curve on the right is the test calibration produced by the Smart Group.

Obviously any person-bound calibration based on the Dumb Group is going to be incomparable with one based on the Smart Group. From the Dumb Group we can only set up percentile ability measures for students who score between 10 and 23. From the Smart Group we can only set them up for students who score between 33 and 46. These two

Figure 1

Person-bound Test Calibration



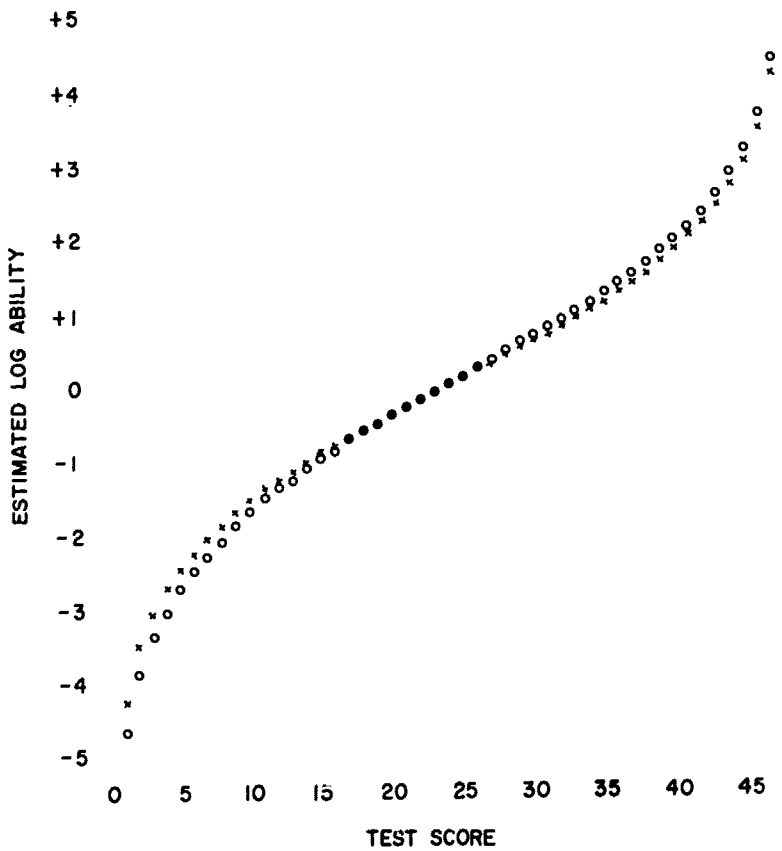
calibrations do not even overlap. And what about all the scores outside the range covered by either group?

Of course Figure 1 describes an exaggerated situation. No one in his right mind would attempt to base a test calibration on two such different groups. But this exaggeration has a purpose. It is aimed at bringing out a treacherous property of person-bound test calibration and providing an acid test for any method which claims to be person-free.

Now let us see how well the new way of test calibration handles this exaggerated situation. I will not burden you with mathematical details.

Figure 2

Person-free Test Calibration



Th.. are covered in the references. (Should you become interested in applying the method, let me know. I have a dandy computer program which does it nicely.) Let us look at the results.

Figure 2 is based on the same data, same test, same students, and the same two calibration curves. But a different method of calibration. As in Figure 1, the x's mark the test calibration based on the Dumb Group. The o's mark the calibration based on the Smart Group. But now, in Figure 2, how different are the two calibration curves?

At this point you may have a question about how calibration curves

work to turn test scores into ability measurements. Each curve represents a conversion table. When a person gets a score on the test, you enter the graph at that score along the bottom, look up vertically to a calibration curve, and then across to the left horizontally to read off his ability. In Figure 2, ability is expressed in logs. If you do not like logs, you can take the antilog and get an ability measure on a ratio scale. This may interest you because then ability is measured on a scale where zero means exactly no ability and for which a regular and meaningful unit can be defined.*

In Figure 1, the calibration curves do not even come close to each other. In Figure 2 they are almost indistinguishable.** Would you say that the difference between the two calibrations in Figure 2 was of practical significance? How much would you care which of these calibration curves you used to make the test a measuring instrument for you? And yet the two groups on which they are based were constructed to make it as hard as possible to achieve person-free test calibration.

One thing that may puzzle you about Figure 2 is the range of test calibration. Either calibration curve provides ability measures for all raw scores on the test from 1 to 47. How can that be done when neither group obtained more than a few of the scores possible?

The answer lies in the item-analysis model on which these calibration curves are based. Remember that this model uses no assumptions about the abilities of the calibration sample. Its only assumption is what happens when any person encounters any item. Out of this assumption it is possible to calibrate a test over its entire range of possible scores even when everyone in the calibration sample happens to get the same score.

That sounds impossible. But it follows directly from this new item-analysis model. The important idea is that even with the same total score, persons differ in those items on which they succeed. When the calibration sample is large, these differences can be used to calibrate the items, and, hence, the test over its entire range of possible scores,

*For a score of 15, the estimated log ability is about -1.0 and the ratio scale ability is about 0.4. A score of 33 indicates a log ability of about $+1.0$ and a ratio scale ability of about 2.7. Thus, a score of 35 indicates about 7 times more ability than a score of 15.

**There is a slight systematic difference. But this reading comprehension test was taken as it stood without any modifications in favor of fitting the item-analysis model. When test items are chosen to conform to the statistical requirements of the model, then no systematic differences between calibrations are discernible.

even though only one score has actually been observed.

Comparing the calibrations shown in Figures 1 and 2, then, we can see the contrast between the present way of doing things—calibration based on the ability distribution of a standardizing sample—and a new way of doing things—calibration that is free from the effects of the ability distribution of the persons used for the calibration. Which do you prefer?*

Item-free Person Measurement

So much for person-free test calibration. Now, how about the companion question? Can ability be measured in a fashion that frees it from dependence on the use of a fixed set of items? Is item-free person measurement possible? If a pool of test items has been calibrated on a common scale, can we use *any* selection we want from that pool to make statistically equivalent ability measurements?

In order to judge whether person measurement can be independent of item selection, we want a situation that will make it as difficult as possible for person measurement to be item-free. For this we will divide the 48 items on the original test into two subtests of 24 items each with *no* items in common between them.

It would be tempting to make these subtests equal in overall easiness. Then they would be parallel forms. But that would be too tame to challenge a scheme for item-free person measurement. Instead, the two subtests will be made as different as possible. The 24 easiest items will be used to make an Easy Test. The 24 hardest items will be used to make a Hard Test. Now, under these circumstances, what is the evidence that ability measurement can be item-free? In other words, what is the evidence that the ability estimates based on the Easy Test are statistically equivalent to those based on the Hard Test?

---

*Even if you use this new way as your basis for calibration, you can still construct all the percentile standardizations you want. Nothing will prevent you from embedding your ability measures in as many sample contexts as you like. But, and this is the vital point, you will not be *bound* by those contexts. You will have an ability measure which is invariant with respect to the peculiarities of the persons used to establish the test calibration. If you were a test manufacturer, you would not have to worry about whether you had obtained the right standardizing samples to suit your customers. Your test would be equally valid for all situations in which the test was appropriate. At the same time, since the calibration was person-free, you would be able to use new data as they came in, to verify and improve item calibration, to add to the item pool, and to document the scope of situations in which the test was functioning properly.

Why do I say statistically equivalent? We know that there are a wide variety of factors at work when a person takes a test. Even knowing a person's ability and an item's easiness will not tell us exactly how he will do on the item. At most we can say what his *chances* are. This uncertainty follows through into his test score. Even if we could give a person the same test twice, wiping all memory of the first exposure from his mind before his second trial, we would not expect him to get the same score both times. We know there will be some variation. This uncertainty is an inevitable part of the situation. It is the error of measurement.

In finding out just how item-free person measurement can be, we must make allowance for this uncertainty. There is no point in asking whether estimates of ability based on the Easy Test are identical with those based on the Hard Test. We know they cannot be. But we can ask whether the two estimates are close enough so that their differences are about what we expect from the uncertainties in the testing situation. Are they close enough in the light of their error of measurement to be considered statistically equivalent?

To answer this question we will examine the test responses of the 976 law students to the 48-item test. The score each student earned on the whole test can be split into a subscore on the Easy Test and a subscore on the Hard Test. This gives each student a pair of independent scores each of which should provide an independent estimate of his reading comprehension ability. In order to convert these scores into ability measures on a common scale, we will calculate calibration curves like the one in Figure 2 for each of the subtests. To do this, we will use item calibrations on a scale common to all 48 items. Then the separate calibration curves for the Easy and Hard tests will convert scores on these different tests into ability estimates on a common scale. If the data fit the item-analysis model, then the independent results from these two different tests should produce statistically equivalent ability estimates.

The data are in Table 1. The upper half of the table is an obvious example of item-bound person measurement. The 976 law students average 6.78 points more on the Easy Test than they do on the Hard one. This problem has been handled in the past by referring such test scores back through a percentile table based on some well-chosen standardizing sample of students who have taken both forms. That is one way to equate two tests which are supposed to measure the same ability. The trouble is that this equation depends on the characteristics

Table 1

Item-free Person Measurement

	Test Score		
	Easy Test	Hard Test	Difference
Mean	17.16	10.38	6.78
Std. Error	0.13	0.14	0.11
Std. Deviation	3.93	4.29	3.30

	Estimated Log Ability			Standardized Difference
	Easy Test	Hard Test	Difference	
Mean	.464	.403	.061	0.003
Std. Error	.032	.028	.024	0.032
Std. Deviation	.997	.868	.749	1.014
Std. Error				0.023

of the sample of persons used to equate the tests. We know that an equation based on one group of persons is not, in general, appropriate for equating measurements made on persons from another group.

Is there a better way to equate tests? Can we go directly from a test score and a person-free calibration of the test items to a measure of ability which does not lean on any particular standardizing sample and which is statistically invariant with respect to those of the calibrated items that are actually used to obtain the score?

The lower half of Table 1 shows how the new approach equates the Easy and Hard tests. We have each person's score on the Easy Test and his score on the Hard Test. For each score we look up the corresponding estimated log ability on calibration curves like the ones in Figure 2. For each pair of scores we obtain a pair of estimated log abilities. They will not be identical. But how do they compare statistically?

The distribution of score differences with a mean of 6.78 and a standard deviation of 3.30 is almost entirely above zero. But the distribution of ability differences with a mean of .061 and a standard devia-

tion of .749 is nicely situated right around zero. On the average, these alternative estimates of ability seem to be aiming at the same thing.

How does the variation around zero compare with what would be expected from errors of measurement alone? To examine this, we will standardize the differences in ability estimates. For each test score there is not only its corresponding ability estimate but also the measurement error that goes with that ability estimate. The difference between the Easy Test and Hard Test ability estimates can be divided by the measurement error of this difference to produce a standardized difference.

It is the distribution of these standardized differences that will show us whether or not the two ability estimates are statistically equivalent. If they are, then this standardized variable should have a mean of zero and a standard deviation of one. That would mean that the only variation observed in ability estimates was of the same magnitude as that expected from the error of measurement in the test. Table 1 shows that, for these 976 students, the standardized differences in ability estimates between the Easy and the Hard tests have a mean of 0.003 and a standard deviation of 1.014. Is that close enough to zero and one?

What does item-free person measurement mean for test constructors and test users? If you can make statistically equivalent person measurements from any selection of items you wish, then all the tricky and difficult problems of equating parallel forms, connecting sequential forms, and relating short and long forms disappear. Incomplete data ceases to be a problem. You can measure a person with whatever items he answers.

Once you have developed a pool of items that conforms to this item-analysis model and once you have calibrated these items, then you are free to make up any tests you wish out of any selection from this item pool. On the basis of these item calibrations alone and without any further recourse to standardizing samples, you can compute a calibration curve or a table of estimated abilities along with their errors of measurement for every possible score on any subtest you want to construct.

All such abilities will be on the same ability scale whatever subset of items they were estimated from. You can measure John on an Easy Test and Jim on a Hard Test and be able to compare their resulting estimated abilities on the same ratio scale. That means you can say how many times more or less able John is than Jim in a precise, quantitative way.

You can measure many children with a short test and a few with a

longer, more precise test and still put all the measures on the same ability scale. Think of how this would expedite screening and selection procedures. The number of items you give a child could depend on how close he comes to the point of decision. Children far away on either side would be quickly detected with a few items. Only children very near the decision point would require longer tests in order to estimate more precisely on which side of the criterion their ability lies.

In general, you would let the required precision, the acceptable error of measurement, determine test length. You would not be bound to any particular predetermined set of items. You could select items from a calibrated pool and compose test forms extemporaneously to suit your measurement needs.* Yet all the measurements made with selections of items from this pool would be located on one scale and used to define whatever norms you or your friends desire. Indeed, since item analyses would be both person- and item-free, it would be easy to construct tests so that *all* new data which came in could be used directly to verify and improve item calibration, to add new items to the item pool, to document the range of persons with whom the test was functioning satisfactorily, and to establish and extend ability norms for whatever groups were being tested.

The Item-analysis Model for Measuring Ability Objectively

By now I hope I have whetted your appetite to know more about the item-analysis model which made these person-free test calibrations and item-free person measurements possible. The measuring model contains just two parameters. One of these belongs to the person and represents the amount of his ability, Z_n. The other belongs to the item and represents the degree of item easiness, E_i. The model combines these two parameters to make a probabilistic statement about what happens when the person tries the item.

Here is the measuring model: The *odds* in favor of success, O_{ni}, are

*The most important criterion for item selection is the magnitude of measurement error. This is minimum when the person being measured has even odds to succeed on the item. That means that we would like to choose items that are just right for the person being measured, items just as easy as the person is able. In individual or computerized testing, where it is possible to choose the next item on the basis of information gathered from the person's performance up to that point, this rule specifies exactly what item to use next.

given by the product of the person's ability, Z_n, and the item's easiness, E_i.*

$$O_{ni} = Z_n E_i$$

This is the same as saying that: The probability P_{ni} that a person with ability Z_n will succeed on an item with easiness E_i is the product $Z_n E_i$ of his ability and the item's easiness divided by one plus this product.**

$$P_{ni} = Z_n E_i / (1 + Z_n E_i)$$

This is the measuring model used to analyze the 48 reading comprehension items on the Law School Admission Test.

What does this simple model say about the scale on which person ability and item easiness are measured? Odds vary from zero to infinity. Since this model gives the odds in favor of success as the product of person ability and item easiness, the natural scale on which to define ability and easiness is one that also varies between zero and infinity.

What does that mean? When a person has no ability, his zero ability will give him zero odds in favor of success no matter what item he tries. With no ability he has no chance of succeeding. On the other hand, if an item has no easiness, then it is infinitely hard and no one can solve it. Measurements made on these scales of ability and easiness have a natural zero.

What about the unit of measurement? Reconsider the product of person ability and item easiness, $Z_n E_i$. There is an indeterminacy in that product. We can multiply ability by any factor we like and not change the product, as long as we divide easiness by the same factor. This shows us that if we want to make measurements, we will have to define a measurement unit.

How can such a unit be defined? One way is to select a special group of items as standard. These items can be chosen on theoretical or normative grounds. They can be chosen because they represent a minimal or optimal level of ability. Once chosen, the combined easiness of these items is set at one. This calibration will then define a person's

*This can equally well be expressed in terms of log odds L_{ni}, log ability X_n and log easiness D_i as

$$L_{ni} = \log O_{ni} = \log Z_n + \log E_i = X_n + D_i.$$

The log odds form brings out the simple linear structure from which this model derives its optimal measuring properties.

**This can equally well be expressed in terms of the logistic function as

$$P_{ni} = 1/(1 + \exp(-(X_n + D_i))).$$

ability as his odds for success on these standard items.

When a person is functioning at about the level of easiness of these items, then his ability is about one. If he is below the level of these items, then his ability is less than one. If, in the course of development or education, he doubles his odds for success, that will mean he has doubled his measured ability. Thus, one way a unit of measurement can be defined is in terms of even odds to succeed on items selected to be standard. The unit of measurement becomes even odds on the standard items.

Another way to define a unit of measurement is in terms of standard persons. These persons can be chosen because they are typical, because they are liminal for some criterion, or because they are the dumbest persons you can find. Now the ability unit is the ability of these standard persons. If you are just at their standard, your ability is one. If your odds to succeed on any item are twice those of a standard person, your ability is two.

In our exploration of what zero means and how to define a unit of measurement, we have uncovered the sense in which measures made with this item-analysis model are on a ratio scale. When one item is twice as easy as another, then any person's odds for success on the easier item are twice his odds for success on the harder one. When one person is twice as able as another, then his odds for success on any item are twice those of the less able person.

Finally, and most important, this simple item-analysis model has a mathematical property that is vital to objectivity in mental measurement. When observations are made in terms of dichotomies like right/wrong, success/failure, it is a mathematical fact that this is the *only* model that leads both to person-free test calibration and to item-free person measurement. When observations are dichotomous, the simple form of this item-analysis model is the *sufficient* and *necessary* condition for objective mental measurement.

Test Construction and the Future of Item Analysis

What bearing does this model for measuring ability objectively have on the construction of mental tests? The model is so simple that those of you who have worried about how to do item analysis may ask: "What about guessing? What about item discrimination? What about the influence of one test item on another?"

It is obvious that in any real testing situation all of these factors play some part. But I prefer to ask: "What do we want to do with them? How big a part do we want guessing, discrimination, and inter-item dependence to play in our measuring instruments?"

We can construct tests in which guessing plays a big part, in which items vary widely in their discrimination, and in which the answer to one item prepares for the next. But do we want to? Not if we aspire to objective mental measurements. If we value objectivity, we must employ our test-conducting ingenuity in the opposite direction.

Most item-analysis models use at least two parameters to describe items. In addition to item easiness, which is part of the simple model presented here, there is also item discrimination. This represents the item's power to magnify or attenuate the extent to which ability is expressed. The discovery of item discrimination was an important step toward understanding how items behave. But as a parameter in the final measuring model it is fatal to objectivity.

If item discrimination is allowed to remain as an active parameter in the measuring model, if gross variation in item discrimination is tolerated in the final pool of test items, then the possibility of person-free test calibration is lost.*

What does this mean for test construction? If we use multiple-choice items, we will devise distractors that make guessing infrequent. When we conduct a pilot study of the characteristics of potential items, we will select items for the final pool that discriminate equally and fit an objective measuring model.

You might complain that this nice advice is impossible to follow. Do not despair. The reading comprehension items on the Law School Admission Test were not constructed for equal discrimination or item independence. They are multiple-choice items with five alternatives. They differ considerably in discrimination, and they are grouped around common paragraphs of text to be read for comprehension. Yet without guessing, without discrimination, and assuming item independence, the simple item-analysis model succeeded quite well, even with these unfit data.

This shows that the measuring model stands up even when one de-

*It may be useful to estimate item discrimination when constructing an item pool in order to bring it under control through item selection. But there are more general statistical tests for whether an item or a set of items fits this simple item-analysis model. Probably these more general tests will turn out to be more generally useful.

parts from its assumptions. We do not have to create a perfect test in order to use the model. That does not mean no further thought need be given to test construction. If we care about building a *science* of mental development, then we must be interested in objective mental measurement. If we are interested in objective mental measurement, then the ideals of no guessing, equal discrimination, and item independence can guide us toward constructing better tests. And the kind of item analysis I have illustrated can transform observations made with these tests into objective mental measurements.

How far have we progressed in the science of mental development since the work of Alfred Binet 60 years ago? I am talking about *science* and not the overwhelming expansion in organization and technique to which our massed presence at this conference testifies. Have you ever wondered why progress is so slow? Something must be wrong. I believe progress will continue to be slow until we find a way to work with measurements which are objective, measurements which remain a property of the person measured regardless of the test items he answers or the company he keeps.

REFERENCES

1. Loevinger, J. Person and population as psychometric concepts. *Psychological Review*, 1965, 72, 143-155.
2. Rasch, G. *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Danish Institute for Educational Research, 1960. Chapters V-VII, X.
3. Rasch, G. On general laws and the meaning of measurement in psychology. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics.* Berkeley: University of California Press, 1961, IV, 321-334.
4. Rasch, G. An individualistic approach to item analysis. In *Readings in mathematical social science.* Edited by Lazarsfeld and Henry. Chicago: Science Research Associates Inc., 1966. Pp. 89-107.
5. Rasch, G. An item analysis which takes individual differences into account. *British Journal of Mathematical and Statistical Psychology.* 1966, 19, Part 1, 49-57.
6. Sitgreaves, R. Review of probabilistic models for some intelligence and attainment tests. *Psychometrika*, 1963, 28, 219-220.

Reformation through Measurement in Secondary Education

PAUL R. LOHNES
State University of New York at Buffalo

Psychometrics has matured as a science to the point where proposals for radical changes in school measurement systems can be drawn from it. Since measurement practices in our schools are not ends in themselves, but exist to further the purposes of the schools, the question arises as to whether radical departures in measurement practices can be expected to change the accomplishments of our schools in desirable ways. My thesis is that there are new approaches to school measurement systems that can promulgate such extensive, desirable changes in our secondary schools as to warrant a hope for reformation through measurement in secondary education.

Schooling is presumed to have its intended effects in the minds of the students who experience it. Nevertheless, among the more concrete traces of a student's path through his schooling are the hieroglyphics inked on his cumulative record and his report cards. These sacred characters memorialize his encounters with an educational measurement system. They represent the ultimate abstraction of the course of his personality development as his teachers have known and guided it. This abstract measurement record is very important to the secondary student because of the meanings others attribute to it and because of the meanings he has been taught to attribute to it himself. My contention is that the measurement record attached to students in most of our secondary schools today is inadequate and harmful; that it involves errors of commission by sponsoring invidious comparisons, burdening teachers, and erecting a barrier between the teacher and the student; that it involves error of omission by ignoring important traits of individual differences in adolescents and providing inadequate interpreta-

tions of the traits it does report.

The purpose of this paper is to sketch in a theory of adolescent personality and a system of educational measurement based on that theory that may point toward new forms of thinking and doing in secondary education. The theory and its measurements have their basis in Project TALENT research.

I assume three particular goals for secondary education: The first is "freeing intelligence through teaching," to borrow Gardner Murphy's formulation (19). Children have to suffer constraints on their intelligence. Primary education tries not to stifle intelligence, but it has to mold it in the process of enculturation. It is especially necessary that children be inducted into the web of conventions comprising the English language. This is a pervasive induction because to know our language the child must know our culture in a very broad way, including knowing its history, its processes, and its values. The child cannot be permitted the freedom not to know, although he should be permitted to wonder and question.

If we want an adult society of free men and women, people need to develop the quality of being free during adolescence with the help of their secondary schooling. The emphasis is on freeing the adolescent's intelligence, not on freeing him in other ways. The school is seen as primarily concerned with the quality of the cognitive functioning of the student. As a member of a family and as a citizen in an industrial democracy, the young adult cannot be free of the bonds of love, loyalties, customs, and laws. He can, however, have a free intelligence with the help of his secondary education.

A person of free intelligence faces predicaments, which David Tiedeman has observed are conditions of human existence, such as the inevitability of death, and of our industrial society, such as the reality of interdependence, which cannot be operated on in a problem-solving mode but must be understood and accepted. There is the predicament of goals, which is that a person has to commit himself to specific goals if he is to achieve anything; yet his free intelligence speculates continually about the alternatives available and treats his commitments as tentative and perhaps temporary. There is the predicament of choices, which involves the realization that to refuse to choose is to refuse to live. "To be understood, a predicament must be experienced; it must be analyzed; . . . it must be practiced in conditions for supervision and discussion" (24).

The second goal is to provide for the insightful study of human pre-

dicaments. Only when it does this can the secondary school help the adolescent to develop the sense of identity that marks the mature person.

Intellectual freedom and personal identity generate personal problems. The self-directed person has to lay out plans and strategies for the long haul and for day-to-day living. These have to be consistent with his personal values and potentials, and then he has to develop his personal abilities to meet the requirements his plans will place upon him. Thus, the third goal is to "cultivate understanding of his personal initiative" in the student (24). The shaping of plans and strategies consistent with values and potentials is assisted by the guidance programs of the secondary school. The development of abilities required by plans and strategies is assisted by the curriculum programs. This third goal has three implications: 1. That guidance and curriculum programs are subordinated to the superordinate educational purposes of freeing intelligence and sponsoring identity; 2. that guidance and curriculum programs cooperate in helping adolescents to transform personal developmental problems into optimal adult adjustments; and 3. that secondary education should sponsor a sense of career in every youth by emphasizing career planning and career adjustments. Girls should be aware that being a housewife and a mother provides a valid career.

If these are goals of particular importance (and I feel they are), what then is the model for secondary education that follows from them? And what are the tasks for an educational measurement system under this model?

Primarily the model specifies the social atmosphere, or morale, of the school. The school in this model is permeated with respect for individual personality. Everyone is encouraged to meet everyone else on a plane of person-to-person transactionalism. This is a community of scholars, bound together by love for each one's own and each other's personality, and by a shared passion for understanding of reality. Teachers and students learn in reciprocity. Learning is motivated not by the press of an extrinsic reward-punishment system but by the freeing in each person of his "deep-seated, persistent need for cognitive orientation to life. . . . The teacher must help the learner to believe in his own individuality and his capacity to learn" (19).

Secondly, the model specifies the foci of educational contents in the school. The primary focus is on human predicaments, from which identity is learned. Here is found the general educational core. The secondary foci are personal plans and strategies. Here is found the

guidance program. The tertiary foci are selected learning goals for individual development according to individual plans and needs. Here is found the individually prescribed curriculum program, as John Flanagan (10) has described it.

The tasks for a measurement system in a school of this model are:

1. To assist the faculty and students to learn and apply a theory of personality that emphasizes the individuality of persons. The problem is that too many teachers view students as a group phenomenon, to be treated by group practices. We need a school measurement system that makes it impossible for teachers to type students. The system would treat each student as a unique person. A teacher's professionalism would encourage her to understand the theory of personality behind the measurement technology she collaborates with. The theory should be sound and compelling so that understanding will create loyalty, and teachers will want to encourage their students to learn and practice this view of personality.
2. The measurement system should create and maintain the school's records of student appraisals in terms of the rubrics of the theory of personality and should be able to provide interpretations of appraisals as prognostic indicators for personally relevant criterion variables. Whereas the first purpose of the system is to sponsor an educational psychology that incorporates the realities of individual differences, the second is to inform students about their potentialities *vis-à-vis* self-posited goals. This task requires that the system incorporate a wide range of predictive validities of the appraisal traits —the results of extensive research into human development. It must also incorporate the statistical procedures for processing answers to specific student's questions from the intersection of the student's appraisal profile and the research findings. As William Cooley has observed (5, 6), the system will have to be computer-based. Only a computer could manage the data retrieval and data analyses required.

What about the origin and outline of a theory of personality that can provide the measurement rubrics and techniques for such a computer measurement system for secondary education? What is required is a theory that provides good descriptions of the status of individuals on educationally relevant variables at various stages of development. For our purposes, the attributions of the theory must be objective and quantitative. The set of variables needs to be simple because the theory will have to be learned and practiced by teachers and students. Such sim-

plicity has two aspects: The number of measurement concepts, or variables, should be small—between 10 and 20. The concepts should already exist in the ordinary language of educational psychology and, so far as possible, of educators in general. The theory should consolidate a provisional doctrine on the description of adolescent personality and a provisional core of measurements for secondary education.

An obvious place to seek such a theory is within the domain of trait and factor psychology. Anne Anastasi's recent book of readings (2) illustrates that the history of this branch of psychology represents one of the finest and fullest chapters in the emergence of the science, and indeed the present problem is an embarrassment of riches. The many theories collected under the generic nomenclature of trait psychology have a good deal in common, but they also encompass significant variations. Each separate theory has sponsored special trait concepts of its own and then devised measurement procedures to operationalize them. The resulting proliferation of tests and inventories has provided researchers with a variety of assessment procedures, but also has disseminated a Babel of concepts. For example, the Project TALENT research to be discussed in this paper is based on exactly 100 observed traits of adolescents.

Fortunately, the theory of factors of behavior-trait intercorrelations and the newly computerized factoring methods provide rationale and technique for digestion of this rich fare. Just how rich the fare is can be illustrated by pointing out that there are 4,950 correlation coefficients describing the interrelationships among the 100 observed traits in the TALENT data. No theory can incorporate 4,950 parameters. The mind boggles before such a theory of mind.

The Value of Factor Analysis

Anne Anastasi (1) has succinctly characterized factor analysis:

> Factor analysis is not a device for discovering basic, immutable units of behavior but a technique for introducing order into a mass of otherwise unmanageable facts.

Truman Kelley (15) put the matter this way:

> There is no search for timeless, spaceless, populationless truth in factor analysis; rather it represents a simple, straightforward problem of description in several dimensions of a definite group functioning in definite manners . . .

Paul R. Lohnes

Factor analysis assists the researcher in his efforts to organize and summarize his data. First and foremost, factor analysis is a heuristic procedure, capable of discovering principles of classification for observations. It is an example of the kind of inductive logic which, when taught to computers, enables artificial intelligence to extend and supplement human intelligence. We need to recognize the heuristic capability of factor analysis, but we also need to perceive clearly that what is discovered by the method are scientific constructs that exist only in the realm of ideas. Cyril Burt (4) expresses this truism:

> A factor is not to be regarded as a simple, isolated, casual entity, much less as an elementary capacity, inherited as such, and capable of spontaneous maturation, regardless of environmental influence . . . A factor is primarily a principle of classification; it is thus not so much a concrete cause as an abstract component.

Just as trait and factor psychology is a generic term for a family of theories, factor analysis is a generic term for a family of methods. Before I could do factor analysis research on the Project TALENT data, I had to make methodological choices. The choices made flowed from allegiance to a particular tradition in trait psychology, for which T. L. Kelley was a leading spokesman.

Truman Lee Kelley (1884-1961) worked at the center stage of educational psychology in America for nearly 50 years, researching, teaching, and writing from the eminences of Columbia, Stanford, and Harvard. His leadership was pervasive, spanning the fields of measurement, statistics, and personality theory. Among his many theoretical and practical accomplishments one commitment stands forth, manifesting itself everywhere in his work, and that is his commitment to the principle of "modes of mental functioning which are independent of other modes" (16). It was his firm and lasting conviction that the "essential traits of mental life" (his title for a 1935 book) would have to be uncorrelated among themselves if they were to have maximum scientific value and practical utility. He argued for this principle repeatedly, worked to develop the necessary methodology of orthogonal factor analysis, and applied the principle consistently in his measurement researches. He held to the principle stubbornly, despite the impossibility of computing large-scale orthogonal solutions with the crude computing machinery of his time, and despite the fact that other measurement psychologists were following the direction taken by L. L. Thurstone with his famous Primary Mental Abilities (23) by resorting

to correlated, or oblique, factor solutions.

In 1928, in *Crossroads in the Mind of Man* (13), Kelley said:

> The advantages of measures of traits which are independent of the other traits involved are so great for all problems of guidance, classification, and education that they are, in truth, at the foundation of a new psychology which the future is to build.

Kelley was an early advocate of the principal-components factoring method. He published a principal-components solution in 1934 for 16 surface traits in a citizenship syndrome (14). Kelley was aware that there is an infinity of orthogonal factor solutions for any correlation matrix, but he argued that the principal components are especially worthy because the major components maximize the extraction of variance from a battery by a subset of factors and produce source traits "in which there are glaring individual differences, not trivial ones" (15).

We discover in Kelley's 1940 presidential address to the Psychometric Society, titled "The Future Psychology of Mental Traits" (17), a compelling set of standards for a factor theory of adolescent personality:

> There are certain fundamental principles which should influence our selection (of derived measures):
>
> The original variables should be wisely chosen and weighted so as to encompass the life situations which it is desired to explain psychologically. The factors comprising the final set should be uncorrelated.
>
> These factors should be ordered for magnitude; this ordering, if the original variables have been wisely chosen and weighted, is also an ordering for importance . . .
>
> The factors comprising the final set should be as stable as possible with changes of age, thus avoiding new factors and new interpretative devices as growth takes place.
>
> As a final practical guide the final factors should be determined with high precision and with low time, administrative and scoring cost.

The obstacle which blocked the engineering of a measurement system on this prescription in Kelley's time was, of course, that there was no machinery capable of handling the monstrous numerical analyses required for the creation and operation of such a system. Not until the

advent of the computer, which came at the end of his life, did the technological revolution he called for occur. We can hope that Kelley finally realized that the road ahead was unblocked.

In order to capitalize on the computer, trait psychologists had to organize support for their research on a new and vast scale. John Flanagan persuaded the U. S. Office of Education to underwrite a national program of research into the distribution, organization, and developmental consequences of abilities and motives of high school youth. In 1960, Project TALENT collected extensive measurement profiles on a probability sample of 440,000 students representing approximately 5 percent of the students in grades 9 through 12 in the nation's secondary schools. The profiles included 60 ability tests and 38 motive scales. In both domains, the measurement instruments represented a good approximation of the state of the art of educational measurement. I have addressed myself to the task of factor analyzing this data in an effort to establish a parsimonious but efficient factor model for it (18).

The a priori value judgments that established the general form of the theory included the decision to have separate factor solutions for the two domains of personality traits—the maximum performance traits (abilities) and the typical performance traits (motives); the decision to have a common solution in each domain span the four years of high school and both sexes; and the determination to have orthogonal factors in each domain. The insistence on orthogonality is going to trouble those who are mindful that most of the major factor theories involve oblique factors. There has been a presumption that uncorrelated factors cannot be satisfactory in the area of interpretability and construct validity. The chief rebuttal is that the factors in both domains produced by this research appear to have strong construct validities and to be unambiguously interpretable. No doubt the simplicity of structures produced by Varimax rotations could be further improved by oblique rotations, so the argument for orthogonality rests finally on practical considerations. The trouble with an oblique structure is that the correlations among the factors require explanation, so the scientist has to generate explanations of explanations.

In the present theory, the locus of orthogonality of the factors of a domain is within a subpopulation composed of students of a single sex and a single grade in high school, although the factor rubrics apply to both sexes and four grades. Sex and grade have been treated as design variables in a linear model. There is a constant effect for sex and a constant effect for grade on each factor for all members of a particular

sex-grade combination. Natural cross-correlations of the ability factors with the motive factors have been studied by canonical correlation procedure.

In the abilities domain, the six main factors out of eleven dimensions derived are three core educational achievements, namely *Verbal Knowledges, English Language,* and *Mathematics;* and three differential aptitudes, namely *Visual Reasoning, Perceptual Speed* and *Accuracy,* and *Memory. Verbal Knowledges* is the chief explanatory concept for 25 different surface traits of specialized knowledges and for the reading comprehension test. Since it is positively correlated to some extent with every one of the 60 ability tests, *Verbal Knowledges* qualifies as a general intelligence source trait, or *g* factor. Spearman said in 1927 that *g* "consists in just that constituent—whatever it may be—which is common to all the abilities" (21). He spoke at that time of the "indifference of the indicator" to emphasize that a measure of *g* can be extracted from any set of maxium performance items for which the performances are mediated by symbol processing. The predominance of special knowledges as primary indicators of this *g* factor reflects Flanagan's experience with Air Force testing programs, in which information tests proved to be the most useful predictors of criterion performances (9).

English Language is a language mechanics ability, the best indicators of which are tests of spelling, capitalization, punctuation, usage, and expression. *Mathematics* is an advanced mathematics and physics ability in which arithmetic computation and arithmetic reasoning do not figure. As shown in the tables, the meaningful factor structure coefficients for the ability factors are all positive, and the structure is simple. The cleanness of the structures in both domains is a tribute to the ingenuity of Henry Kaiser's Varimax rotation scheme (12), which was used to create them. Since one cannot compute uncorrelated linear functions without the use of beta weights of mixed signs, the functions of ability revealed by the factor score matrix are not as neat as the view given by the factor pattern.

In the motives domain, 11 dimensions were extracted also, and the main factors have been named *Conformity Needs, Scholasticism, Activity Level,* and four interest factors—*Business, Outdoors and Shop, Cultural,* and *Science. Conformity Needs* may be seen as a measure of the extent to which a youngster subscribes to the middle class mores of our society, or of what Edwards calls "Social Desirability" (8). *Scholasticism* represents a measure of the student's devotion to academic pursuits. Somewhat surprisingly, the canonical correlations

Table 1

60 Abilities Domain Variables from Project TALENT

	Mnemonic	*Code*	*Name of Test*		*Mnemonic*	*Code*	*Name of Test*
1	SCR	R-101	Screening	35	THR	R-150	Theater and Ballet
2	VOC	R-102	Vocabulary				
3	LIT	R-103	Literature	36	FDS	R-151	Foods
4	MUS	R-104	Music	37	MIS	R-152	Miscellaneous
5	SST	R-105	Social Studies	38	MMS	R-211	Memory for Sentences
6	MAT	R-106	Mathematics				
7	PHY	R-107	Physical Sciences	39	MMW	R-212	Memory for Words
8	BIO	R-108	Biological Sciences	40	DSW	R-220	Disguised Words
9	SCA	R-109	Scientific Attitude	41	SPL	R-231	Spelling
10	AER	R-110	Aeronautics and Space	42	CAP	R-232	Capitalization
				43	PNC	R-233	Punctuation
11	ELE	R-111	Electricity and Electronics	44	USG	R-234	English Usage
				45	EXP	R-235	Effective Expression
12	MEC	R-112	Mechanics				
13	FAR	R-113	Farming	46	WDF	R-240	Word Functions in Sentences
14	HEC	R-114	Home Economics				
15	SPO	R-115	Sports	47	RDG	R-250	Reading Comprehension
16	ART	R-131	Art				
17	LAW	R-132	Law	48	CRE	R-260	Creativity
18	HEA	R-133	Health	49	MCR	R-270	Mechanical Reasoning
19	ENG	R-134	Engineering				
20	ARH	R-135	Architecture	50	VS2	R-281	Visualization in Two Dimensions
21	JUR	R-136	Journalism				
22	FOT	R-137	Foreign Travel				
23	MIL	R-138	Military	51	VS3	R-282	Visualization in Three Dimensions
24	ACC	R-139	Accounting				
25	PRK	R-140	Practical Knowledge	52	ABS	R-290	Abstract Reasoning
26	CLE	R-141	Clerical				
27	BIB	R-142	Bible	53	ARR	R-311	Arithmetic Reasoning
28	COL	R-143	Colors				
29	ETI	R-144	Etiquette	54	MA9	R-312	Introductory Mathematics
30	HUN	R-145	Hunting				
31	FIS	R-146	Fishing	55	ADV	R-333	Advanced Mathematics
32	OUT	R-147	Outdoor Activities (other)	56	ARC	R-410	Arithmetic Computation
33	PHO	R-148	Photography	57	TBL	R-420	Table Reading
34	GAM	R-149	Games (sedentary)	58	CLR	R-430	Clerical Checking
				59	OBJ	R-440	Object Inspection
				60	PRF	A-500	Preferences

Table 2

Abilities Domain Factors

Mnemonic	*Factor Name*	*Variance Extracted*
VKN	Verbal Knowledges	18.7%
GRD	Grade	7.8%
ENG	English Language	6.6%
SEX	Sex	5.7%
VIS	Visual Reasoning	5.3%
MAT	Mathematics	4.1%
PSA	Perceptual Speed and Accuracy	3.6%
SCR	Screening	3.3%
H-F	Hunting-Fishing	2.2%
MEM	Memory	2.1%
COL	Color, Foods	1.9%
ETI	Etiquette	1.6%
GAM	Games	1.5%

(13 factors extract 64.6% of variance)

between domains revealed little common variance between the ability factors and the motive factors. There is only about 10 percent of redundant variance in either set of factors, given the other set. Most of the redundancy is due to the relationship between academic achievement and academic orientation. The factors of the two domains provide substantially autonomous subsystems of personality constructs in the descriptive theory.

These factors are proposed as suitable variables for a computerized measurement system for secondary schools. They have the required simplicity in number and in ordinary language accessibility. What remains to be shown is their predictive validities as prognostic indicators of future development. William Cooley and I have just sent to press a Project TALENT monograph which documents the validities of these factors for several important educational and vocational development criteria from follow-ups of subjects one and five years out of high school (6). High school curriculum, type of post-high-school institution, and college majors are shown to be predictable. We have organized a career development tree structure that spans the years from elementary school to young adulthood (Figure 1 on page 117). This Career Development

Table 3

Abilities Domain Variable-factor Correlations ≥ .35

Test	VKN	GRD	ENG	SEX	VIS	MAT	PSA	SCR	H-F	MEM	COL	ETI	GAM	h^2	R^2
SCR								61						64	40
VOC	66													79	78
LIT	69	42												76	73
MUS	65													63	59
SST	70													77	76
MAT	45					62								82	75
PHY	54					42								74	71
BIO	51													63	56
SCA	47													52	49
AER	50			42										63	57
ELE	36			44										69	64
MEC				52				38						74	69
FAR	36							47						65	50
HEC				−52										66	59
SPO	48			39										57	55
ART	72													68	63
LAW	61	35												58	53
HEA	56													60	56
ENG	39													48	42
ARH	53													40	33
JUR	58													49	45
FOT	68													57	50
MIL	59													51	38
ACC	54	39												54	53
PRK	47													58	46
CLE		53												51	48
BIB	63													60	45
COL											66			65	27
ETI												71		79	21
HUN				43					58					59	34
FIS									77					74	23
OUT	50													55	49
PHO	41													40	30
GAM	41												46	53	29
THR	65													64	60
FDS	46										51			59	35
MIS	63													56	52
MMS										83				86	20
MMW										50				57	38
DSW	46		40											65	58
SPL			58											67	56
CAP			62											59	43
PNC	38		60											75	69
USG	36		59											62	54
EXP			53											57	46
WDF	40		42											66	58
RDG	65	35	39											81	79
CRE	46				41									57	53
MCR				44	59									73	66
VS2					63									57	36
VS3					71									66	49
ABS					57									64	54
ARR	41		39											66	63
MA9	39		36			61								79	73
ADV						71								69	46
ARC			46				36							67	54
TBL							71							59	36
CLR							76							65	38
OBJ							67							62	35
PRF							56	35						64	18

Table 4

38 Motives Domain Variables from Project TALENT

	Mnemonic	Code	Name of Scale
1	MEM	A-001	Memberships
2	LEA	A-002	Leadership Roles
3	HOB	A-003	Hobbies
4	WOR	A-004	Work
5	SOC	A-005	Social
6	REA	A-006	Reading
7	STU	A-007	Studying
8	CUR	A-008	Curriculum
9	COU	A-009	Courses
10	GRA	A-010	Grades
11	GUI	A-011	Guidance
12	NSO	R-601	Sociability
13	NSS	R-602	Social Sensitivity
14	NIM	R-603	Impulsiveness
15	NVI	R-604	Vigor
16	NCA	R-605	Calmness
17	NTI	R-606	Tidiness
18	NCU	R-607	Culture
19	NLE	R-608	Leadership
20	NSC	R-609	Self-confidence
21	NMP	R-610	Mature Personality
22	IPS	P-701	Physical Science, Engineering, Mathematics
23	IBS	P-702	Biological Science, Medicine
24	IPU	P-703	Public Service
25	ILL	P-704	Literary, Linguistic
26	ISS	P-705	Social Service
27	IAR	P-706	Artistic
28	IMU	P-707	Musical
29	ISP	P-708	Sports
30	IHF	P-709	Hunting, Fishing
31	IBM	P-710	Business Management
32	ISA	P-711	Sales
33	ICO	P-712	Computation
34	IOW	P-713	Office Work
35	IMT	P-714	Mechanical, Technical
36	IST	P-715	Skilled Trades
37	IFA	P-716	Farming
38	ILA	P-717	Labor

Tree is an attempt to consolidate in one model features of the developmental theories of Eli Ginzberg (11) and of Donald Super (22) and the vocational classification theory of Anne Roe (20). Transitions within the paths of the tree are found to be related to factor profiles. In the aggregate, career plan changes are found to follow a probability law which states that people tend to select new career objectives that place them in groups they resemble psychometrically more than they resemble the groups they migrate away from. The psychometric taxonomy of vocations provided by the 12 branch tips of the tree represents a useful organization of the world of work for guidance purposes. The categories are discriminated by the profiles describing the average person in each category.

Table 5

Motives Domain Factors

Mnemonic	*Factor Name*	*Variance Extracted*
CON	Conformity Needs	11.1%
SEX	Sex	9.1%
BUS	Business Interests	8.7%
OUT	Outdoors, Shop Interests	6.8%
SCH	Scholasticism	6.6%
CUL	Cultural Interests	5.8%
SCI	Science Interests	4.3%
GRD	Grade	4.2%
ACT	Activity Level	4.0%
LEA	Leadership	3.1%
IMP	Impulsion	2.8%
SOC	Sociability	2.8%
INT	Introspection	2.4%

(13 factors extract 71.5% of variance)

This career tree and vocational taxonomy could be taught to secondary school students as part of a curriculum on the world of work. Then the computer measurement system could provide personal interpretations of factor score profiles of individual students *vis-à-vis* their personal planning questions, phrased in terms of the categories of this theory of careers. These outputs from the measurement system would be prognoses of potential personal futures, expressed in the language of probabilities. They would help the student, perhaps in a counseling context, to make educational and vocational decisions. The student would be helped to locate possible goals which have high probability for him. If he holds to a goal that is shown to have a low probability for him, he will be able to see which of his traits need to be modified to increase the probability. The student is helped to see the real probabilities of various outcomes *for him;* he sees that his future is not determined, since probabilities are a joint function of his goals and his attributes, and both are subject to deliberate change.

Table 6

Motives Domain Variable-factor Correlations ≥ .35

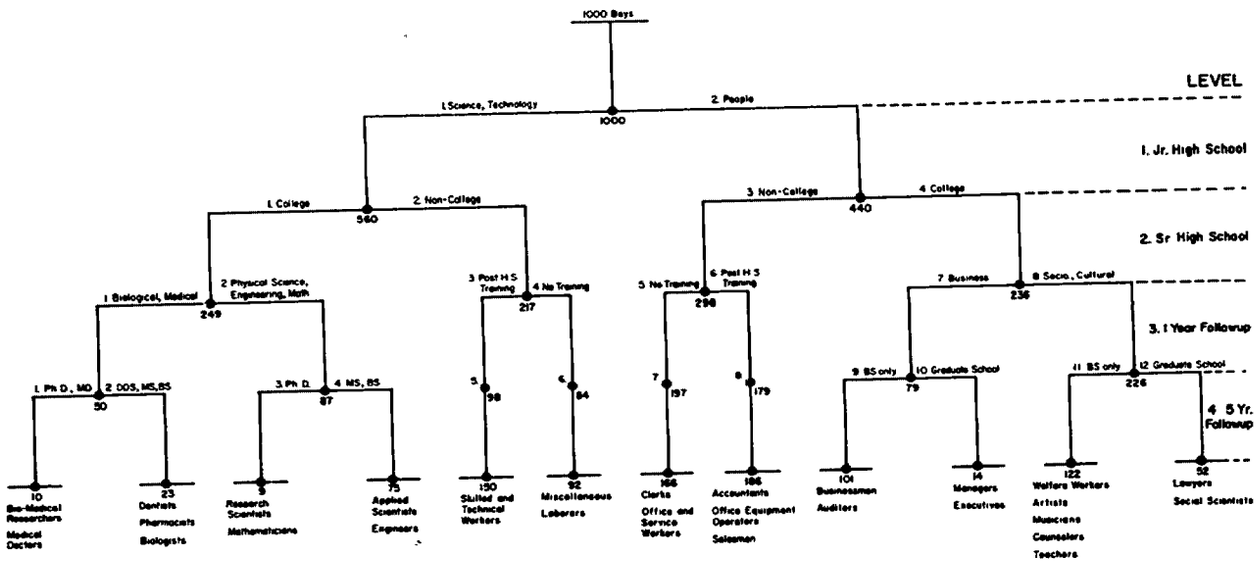
Test	CON	SEX	BUS	OUT	SCH	CUL	SCI	GRD	ACT	LEA	IMP	SOC	INT	h^2	R^2
MEM									60					61	31
LEA										83				75	17
HOB									62					68	44
WOR									71					64	29
SOC												62		66	26
REA					39								55	66	25
STU					72									74	52
CUR					70									62	35
COU					53			44						56	40
GRA					75									66	41
GUI					55									54	39
NSO	63											43		68	48
NSS	72													66	56
NIM											87			83	16
NVI	67													61	45
NCA	74													66	52
NTI	75													68	53
NCU	72													70	58
NLE	51									44				61	39
NSC	45												66	76	30
NMP	78													75	64
IPS		47					62							82	77
IBS							74							75	56
IPU			51				37							64	55
ILL			39			68								82	72
ISS		—49	46			35								65	63
IAR						70								70	55
IMU						77								70	44
ISP		35		50										68	50
IHF		50		61										72	58
IBM			74											78	71
ISA			74											68	58
ICO			79											73	62
IOW		—55	62											74	67
IMT		63		51										80	83
IST		35	45	67										84	81
IFA				77										73	55
ILA			45	61										79	68

Conclusion

In this paper, I have proposed a number of fundamental changes. First, I proposed that grading of students by teachers be eliminated. Periodic appraisals of students would be conducted by a system of uniform measurements. Grading violates transactionalism by forcing the teacher to exhibit disrespect for the student's individuality. Grading not only denies the uniqueness of the student; it lends itself to the substitution of extrinsic for intrinsic learning motivation. Removing the grading assignment from teachers to some other agent such as the Regents or a

Figure 1

Career Development Tree from Project TALENT



computer will not solve the problem if the student perceives a continuing emphasis on invidious comparisons such as class rankings. He will still blame the teacher for being a party to such treatment of him. We require a measurement system in our schools that avoids ritualistic and punishing comparisons, emphasizing instead the personal values of more and better information about oneself that individual students can realize as a benefit of the system. This can be accomplished by emphasizing the role of measurements as prognostic indicators, validated by follow-up researches. This emphasis would encourage the student to be interested in his measurement profile for what it can reveal to him about his personal purposes and potentialities. Even so, self-referring information is necessarily punishing at times, and how much better it would be if its source were an objective measurement system rather than the teacher's judgments. The teacher should lead and help students, not judge them.

Second, it is proposed that teachers and students learn new rubrics for appraising personal development in adolescence. These are the rubrics of independent factors of ability and motive. The major ability factors proposed are:

1. Verbal Knowledges
2. English Language
3. Mathematics
4. Visual Reasoning
5. Perceptual Speed and Accuracy
6. Memory

The major motive factors proposed are:

1. Conformity Needs
2. Scholasticism
3. Business Interests
4. Outdoors and Shop Interests
5. Cultural Interests
6. Science Interests

These categories are the main constructs of a descriptive theory of adolescent personality that emphasizes the uniqueness of each person, the prognosticated potentials of each person, and the malleability of personality as a function of personal initiatives. The theory is a network of research-based relationships between these variables and important environmental and adjustment variables.

Since schools would keep the cumulative records of students in terms of these rubrics, they would become elements of the ordinary language of education. Proposing these rubrics does not prescribe the Project TALENT test battery as the only solution to the measurement instrumentation requirement. Research has already shown that these factors can be successfully regressed on various measurement batteries to provide a variety of bases for scoring estimates of the factors. There remains a need for development of forms of scaling these variables that will be especially sensitive to changes in status over relatively short periods of time. Benjamin Bloom has made it clear that grades assigned by teachers are not measures of change, and that students need appraisals of changes they develop in their trait characteristics (3). It is proposed that the Project TALENT longitudinal validities for the factors, representing the nation's largest investment in follow-up educametric research, should be incorporated in the computer measurement system. This does not exclude the incorporation of research findings from other national or local follow-up studies.

I started with three goals for secondary education and have described a measurement theory and system that I think would be able to cope with the tasks posed for it in schools pursuing those goals. The first goal is that the school should be a place where the student frees his intelligence. Learning the theory of personality sponsored by this measurement approach would equip the student with conceptual tools to enhance his understanding of himself and his peers. If the teaching is conducted in a discovery mode, this learning could be one of the experiences that free intelligence. Meanwhile, the purging of teacher grading in favor of the new measurement system would make transactional teaching more of a possibility in our schools. The other goals are that the school should sponsor personal identity and personal initiatives. The prognostic outputs of the measurement system would stimulate and assist the student in understanding his predicaments and planning his initiatives. The goals imply that guidance and curriculum services should be coordinated to provide individualized learning sequences in accordance with each student's career concepts. This measurement system would support the exploration of potential careers and the learnings required to qualify for them.

If we want to have secondary schools in which adolescents develop into self-directing adults with free intelligence and responsible orientations toward productive careers, we have to provide a theory and practice of educational measurement conducive to these goals.

REFERENCES

1. Anastasi, Anne. *Differential psychology, third edition.* New York: The Macmillan Company, 1958.
2. Anastasi, Anne. *Individual differences.* New York: John Wiley & Sons, Inc., 1965.
3. Bloom, B. S. *Stability and change in human characteristics.* New York: John Wiley & Sons, Inc., 1964.
4. Burt, C. The structure of the mind: a review of the results of factor analysis. *British Journal of Educational Psychology,* 1949, 100-111, 176-199.
5. Cooley, W. W. A computer-measurement system for guidance. *Harvard Educational Review,* 1964, 559-572.
6. Cooley, W. W. and Lohnes, P. R. Implications for guidance. In J. C. Flanagan, *et al., Project* TALENT *one-year follow-up studies.* Pittsburgh: University of Pittsburgh, 1966. Pp. 225-234.
7. Cooley, W. W and Lohnes, P. R. *Predicting adult development.* Palo Alto: American Institutes for Research, 1967.
8. Edwards, A. L. Social desirability and performance on the MMPI. *Psychometrika,* 1964, 295-308.
9. Flanagan, J. C. *The aviation psychology program in the Army Air Forces, report no. 1.* Washington, D.C.: U.S. Government Printing Office, 1948.
10. Flanagan, J. C. *Developing a functioning model of an educational system for the '70's.* Palo Alto: American Institutes for Research, 1967.
11. Ginzberg, E., Ginsburg, S. W., Axelrad, S., Herma, J. L. *Occupational choice: an approach to a general theory.* New York: Columbia University Press, 1951.
12. Kaiser, H. F. The varimax criterion for analytic rotation in factor analysis. *Psychometrika,* 1958, 187-200.
13. Kelley, T. L. *Crossroads in the mind of man.* California: Stanford University Press, 1928.
14. Kelley, T. L. and Krey, A. C. *Tests and measurements in the social sciences.* New York: Charles Scribner's Sons, 1934.

15. Kelley, T. L. Comment on Wilson and Worcester's "Note on Factor Analysis." *Psychometrika*, 1940, 117-120.

16. Kelley, T. L. *Talents and tasks.* Cambridge: Harvard University Press, 1940.

17. Kelley, T. L. The future psychology of mental traits. *Psychometrika*, 1940, 1-15.

18. Lohnes, P. R. *Measuring adolescent personality.* Pittsburgh: University of Pittsburgh, 1966.

19. Murphy, G. *Freeing intelligence through teaching.* New York: Harper Brothers, 1961.

20. Roe, Anne. *The psychology of occupations.* New York: John Wiley & Sons, Inc., 1956.

21. Spearman, C. *The abilities of man: their nature and measurement.* New York: Macmillan, 1927.

22. Super, D. E., Crites, J. O., Hummel, R. C., Moser, H. P., Overstreet, P. L., Warnath, C. F. *Vocational development: a framework for research.* New York: Teachers College, 1957.

23. Thurstone, L. L. *Primary mental abilities.* Chicago: University of Chicago Press, 1938.

24. Tiedeman, D.V. Predicament, problem, and psychology: the case for paradox in life and counseling psychology. *Journal of Counseling Psychology*, 1967, 1-8.

Surveys Undertaken by the Scottish Council for Research in Education

DAVID A. WALKER
The Scottish Council for Research in Education

Scotland is a relatively small country with a population of just over five million, of whom about 900,000 are in full-time attendance at school. For the purposes of a survey of a year group, this is a very convenient size. The Scottish educational system has the additional advantage that 95 percent of the pupils (as we describe them) attend public schools, which are administered by the Scottish Education Department and the 35 local education authorities. Denominational schools, mostly Roman Catholic, form part of this system. Two percent of the pupils are in grant-aided schools and two percent in independent schools.

Scotland has also the advantage of an interest in educational research going back over many years. The Scottish Council for Research in Education was instituted 40 years ago. When the suggestion was made in 1931 that there should be a Scottish Mental Survey, the climate was favorable.

The 1932 Scottish Mental Survey

The idea of a mental survey arose from inquiries into the incidence of mental defect, which had been conducted in England and Wales. The original suggestion was that the Council should determine what proportion of the age groups forming the school population should be classed as "mentally deficient," the term then in use. The committee appointed to prosecute the inquiry soon recognized that its objects could be achieved and the value of the investigation enhanced if a complete cross section of the community were studied. The cross section chosen was

the year group covering the older ten-year-olds and the younger eleven-year-olds. The testing date selected was in June 1932, and the population was defined as those children born in 1921. It was not possible to obtain an exact count of them, since in those days returns of numbers were not so complete as they are now, but the population was estimated to number about 96,000. The committee decided to attempt to test all of them, and the cooperation of the authorities administering education, including those controlling the grant-aided and the independent schools, was readily given.

Two tests were used, the first being of the now-familiar group-test type including 76 verbal items and 9 pictorial items. The second was the 1916 Stanford-Binet Scale, an individual test. The group tests were taken by 87,500 children—*i.e.*, about 90 percent of the age group, the losses being largely attributable to normal absence through illness. The test booklets were marked and checked by their teachers, who were given detailed instructions. The Stanford-Binet test was taken by 1,000 pupils in the same age group, selected in a pseudo-random fashion, and was administered by trained testers. The scores and quotients made by this sample in the two tests provided a link between the group test scores and the intelligence quotients of the Binet Scale.

The results of the survey indicated that the average IQ of the Scottish boys was 100 and that of the girls slightly under 100, showing agreement with Terman's standardization. But the standard deviation of the boys' quotients was 17 points of IQ and that of the girls 16, both substantially higher than the figure of 13 obtained by Terman. It was also shown that the proportion of children in the age group who had very low verbal intelligence was higher than previously supposed.

In its report (1), the committee emphasized the wide scatter of scores that had been found and presented its findings as a contribution to the study of the intricate problems which confront a democracy in its endeavor to organize educational opportunity suited to the widely varying needs of the younger generation. What the committee had also established was that a survey of this type was a practical proposition once the cooperation of teachers and administrators had been obtained and an efficient organization set up to handle the data. It had also provided a record of the test scores of a complete age group, probably the first of its kind in any country in the world.

The Macmeeken Survey (1936-37)

In my reference above to the thousand pupils who, in 1932, took both the group and the individual tests, I used the term "pseudo-random sample." The sample was not a truly random one because trained testers were not equally available in the various areas of the country. The results showed that some areas were not adequately represented and that the sample was definitely superior to the rest of the population as far as scores in the group test were concerned.

The Council therefore agreed to undertake the individual testing of a truly random sample of the children in Scotland who were born in 1926. The sample was defined as those born on February 1st, May 1st, August 1st or November 1st in that year, and the task of administering the test was entrusted to one person, Miss A. M. Macmeeken. She began the task in September 1935 and completed it in November 1937.

The names and schools of the selected children were ascertained from returns supplied by all the schools in Scotland having pupils in the appropriate age range. Of the 874 so traced, 873 were tested. One boy, the son of an Irish laborer in Glasgow, had disappeared with his family before the tester reached his school.

Miss Macmeeken (2) found that the average IQ of the boys was 100.5 and that of the girls 99.7, the difference not being statistically significant. The standard deviation of the quotients was again found to be higher than Terman had first found. The survey had therefore verified the results obtained in the 1932 Survey.

In addition to administering the Binet test, Miss Macmeeken, aided part of the time by an assistant, gave a battery of performance tests to each child in the sample. The interesting finding from this auxiliary survey was that there were marked differences between the sexes in these tests—the boys being superior—and the experimenter raised the question whether the equality shown in the Binet test was merely a confirmation of the success of the constructors of the scale in eliminating sex differences.

The 1947 Scottish Mental Survey

Several workers in the field had observed that children in large families tended to have lower scores than those made by children in smaller families. This gave rise to a fear that the national level of intelligence might be falling. The most straightforward way to obtain reliable information on both of these points was to repeat the Scottish 1932 Survey

using the same test.

The Council had intended to repeat the 1932 Survey after an interval of about 25 years, and the material obtained in the survey had been stored with this in view. Discussions in a Population Investigation Committee, which was associated with a Royal Commission on Population sitting in 1945, led to the request that the new survey be held in 1947, 15 years after the first, although there was some doubt whether a 15-year interval was sufficiently long to allow a trend to show.

The main variables for the new survey were test score and size of family, but the Council, in agreeing to conduct it, decided to extend the inquiry to include sociological variables, such as father's occupation, housing conditions, and migration (4, 5).

Once again there was excellent cooperation from all quarters. Of the 76,330 children born in Scotland in 1936 (the year group chosen for the population), survey records were obtained for 75,221, while 70,805 (93 percent of the age group) took the group test, 4,406 being absent from school on the day it was given. The sample selected for the individual test, which was again used to calibrate the group test, numbered 1,230, and 1,215 (99 percent) of these were given form L of the Terman-Merrill revision of the Binet test.

This sample consisted of all children born on the first day of alternate months in the year, beginning with February. A larger sample, which consisted of all children born on the first three days of each month, provided more extensive sociological data. The number in this sample was 7,380, and the home of each of these pupils was visited to obtain the required information. In both samples it was found that the selected children were a fair representation of the population on each variable for which there were data for both sample and population.

The analysis of the data established the negative association between measured intelligence and family size, both for group-test scores and for Terman-Merrill IQs. But it was found that average group test scores had *risen* by a small but statistically significant amount in the interval between 1932 and 1947. The average IQ had also risen but by a non-significant amount.

The survey (3) had therefore provided answers to the two questions which had been posed, but had answered them in so paradoxical a fashion as to raise others. This we all know to be a feature of educational research: More questions may be raised than answers given.

The discussion of the findings has led research workers to probe more deeply into the basic assumptions underlying the administration

of tests of this type and the interpretation of their scores. We are a little clearer now than we were 20 years ago about the terms we use—"intelligence," "intelligence tests"—and on the effects of social and cultural environment on test scores. Further research has provided what may well be the explanation for the paradox.

The 1947 Survey also provided a baseline for a follow-up study (6) which continued for 17 years (the final report is now on press). The group selected for the follow-up was the sample of 1,215, for most of whom there were available Terman-Merrill IQs and sociological data in addition to group test scores. The sample covered the whole range of ability, and in this respect the study differed from others which have been reported. The Council has been fortunate in being able to keep in touch, directly or indirectly, with 92 percent of the sample members.

Another by-product of the survey has been the testing of the younger sibs of the random sample members, as each sib reached the age at which the sample member was tested. Thus the survey not only provided the information originally sought, but generated a number of additional projects (7).

The Scottish Scholastic Survey 1953

In working over the data provided by the Mental Surveys, the investigators had frequently been reminded of the lack of national records of scholastic attainments. It was therefore suggested that the Council should conduct a further survey of approximately the same age group as that previously tested but using on this occasion tests of attainment in arithmetic and English. It was hoped that the survey would indicate the relative educational standards of urban and rural schools and of different sizes of school and would produce norms which would enable teachers to assess, on the basis of national standards, the attainments of their classes or pupils in the age range.

The Council agreed to undertake this task, making it clear that the results would not be published in a form enabling comparisons to be made between individual schools or between education authorities.

The tests were constructed by panels of teachers assisted by experts in test construction. There were four tests assessing attainments in mechanical arithmetic, arithmetical reasoning, English usage, and in English comprehension. They were of the objective type so that the marking could be done by the teachers. This task of marking was willingly undertaken by the teachers who were naturally interested in the

degree of success of their pupils.

Since the survey was intended to serve so many purposes, it was thought advisable to test a whole year group, which was defined roughly as the ten-year-olds. The tests were to be given in June 1953 and the population was defined as those children born between July 1st, 1942 and June 30th, 1943 and attending public, grant-aided, or independent schools. They numbered about 76,000 and of these more than 72,000 took the whole battery of tests—a response rate of 95 percent.

The analysis of the results showed that the differences in achievement between those who lived in cities, in large towns, in small towns, and in other areas were slight. The various regions of the country, ranging from the northern isles to the central industrial belt, also showed no systematic differences; those that were above average in one test were usually below average in another. Pupils in smaller schools attained much the same standards as those in larger schools. The one-teacher schools were superior to those slightly larger in size and equal in standard to the largest schools. Pupils in larger classes tended to make higher scores than those in small classes, a result which requires careful examination before conclusions are drawn. There were marked sex differences in three of the tests. In the arithmetical reasoning test, boys were superior; in both of the English tests, the girls were superior.

The committee supervising the project examined a substantial number of the completed scripts. These were drawn at random, and the members made a careful analysis of the types of error that had been made and endeavored to identify the processes which had given pupils most difficulty. The report of the survey (8) included these findings along with recommendations to teachers of the age group.

From the data of the survey it was also possible to establish norms. However, the report, which contained the full text of the tests, had been delayed and it did not appear until the autumn of 1963. By that time the Council had agreed to repeat the survey, and the norms became obsolete.

The Scottish Scholastic Survey 1963

The purpose of this survey was to ascertain what changes in standards of attainment had occurred in the 10 years between 1953 and 1963. Researches in the fields of mental testing and of reading had shown that test scores had risen in the post-war period, and the possession of the unpublished tests used in 1953 made it relatively easy to plan a

survey which would assess the changes in the attainments of Scottish ten-year-olds. Since comparisons of subgroups were not envisaged, although in fact some comparisons of this type were made, the Council decided to test only a sample of the whole population of 82,000 pupils in 2,600 schools.

The sampling design chosen was a stratified cluster sample, with the school as primary sampling unit and a sampling fraction of 1 in 15 for schools, but no sub-sampling with schools. The stratification was by type of area (cities, large towns, small towns, other areas) and by size of school. The sample numbered 5,209 pupils who were in 169 schools.

As has been indicated, the tests were those of the 1953 survey, printed in the same format and with the same instructions. The response rate was 96 percent, the missing four percent being the normal absence rate at that time of year. Once again the committee examined scripts drawn at random, concentrating on this occasion on items which had shown large or small changes in difficulty level in the interval between the two surveys.

The main finding was that the performance of the 1963 group was markedly superior to that of the 1953 group. The improvement in each of the four tests was roughly equal to the progress made by an average ten-year-old pupil in six months. The gains in score were made by pupils at all levels of ability, by boys and girls to the same extent, in all regions of the country and in all sizes of school. They were spread over nearly all of the items in the tests, and could be attributed partly to greater speed in responses and partly to greater accuracy where responses had been made. In the arithmetic tests, pupils seemed to have greater familiarity with tables of capacity, weights and measures, and money. In English, the tests showed the pupils to be reading with more skill and becoming more independent in their thinking about what they read.

One surprising finding which required further investigation was the high intra-class correlation (about 0.3) within schools. This corresponds to a design effect of about 6—*i.e.*, the 5,209 pupils produced the same precision as a simple random sample of about 800 drawn from the whole population. The correlation and design effect are higher than those usually found in such surveys. The size of the design effect is partly due to the decision not to sub-sample within schools. This decision was taken primarily on administrative grounds and was probably the correct decision in spite of the apparent inefficiency resulting from it.

From the new data fresh norms were prepared and these are included in the report of the survey (10), which is on press. The presentation of the new norms will help to impress Scottish teachers with the need to use up-to-date norms for scholastic tests. This point is stressed in the American Psychological Association's recent booklet, *Standards for Educational and Psychological Tests and Manuals,* but is still unfamiliar to some Scottish teachers.

Scottish Standardization of the Wechsler Intelligence Scale for Children (1961 and 1962)

I have included this as a survey because in execution and outcome it conformed to the patterns of Council surveys. The aim was to prepare a Scottish standardization of the Wechsler test, which had proved useful to Scottish psychologists.

The population was defined as all children between the ages of 5 and 15 attending school, whether public, grant-aided, or independent. The sample was selected as those children in primary schools born on January 25th, and those in secondary schools born on January 31st. The testing date was changed for secondary pupils to prevent a child being tested twice, since the testing was spread over a number of months during which a pupil might transfer from primary to secondary education. The names of the children were supplied by the Directors of Education of the various areas and by the heads of the grant-aided and the independent schools. The sample provided about 200 children in each year group.

After consultation with Dr. Wechsler and the Psychological Corporation, a few minor changes were made in the items to render them more suitable for Scottish children. The testing was undertaken largely by educational psychologists in the employment of local authorities, aided by members of psychology staffs in universities and colleges and in a few cases by students who were given a period of training and supervised practice. In this way almost all children in the age ranges 6-11 and 13-14 were tested. There were, however, gaps among those aged 5, 12, and 15.

From the data obtained, Scottish norms were established and made available to Scottish psychologists. From the survey point of view, there were several interesting results. The Scottish and American norms were found to be in fairly close agreement: There were differences in standards in the tests making up the scale, but not over the scale as a whole.

It was made clear that there was a need for rigorous observance of standard proced re in administering and scoring the scale if a reliable assessment of an intelligence quotient was to be obtained.

The execution of this project required several years of work both in the field and in the laboratory, but it is thought to have been well worthwhile. A description of the organization of the survey has been published (9), but the manual dealing with sc_ ing is available only to users of the scale.

Assessment for Higher Education (1962)

A survey of a very different type was that undertaken by the Council in 1962 as part of a study of measures likely to predict success in higher education—*i.e.*, ' universities, colleges of education, and technical colleges. The population in this case was the group of students who were attempting the examinations set at that time by the Scottish Education Department. The attainment of sufficiently high scores on the required examinations gave admittance to institutions providing higher education. The survey was restricted to those attending public and grant-aided schools; they numbered about 12,000 and were in 221 schools.

The sample was defined as the whole population. Three schools refused to cooperate, and the response rate was about 98 percent; but there are unavoidable gaps in the data for some of these students.

The tests consisted partly of the examinations set by the Scottish Education Department, which granted the Council access to the scores. These examinations have the virtue of being national so that the difficulties of equating standards of different examination boards have been avoided. Through the kindness and cooperation of the College Entrance Examination Board and Educational Testing Service, each student took the Board's Scholastic Aptitude Test, and scores on the verbal and mathematical sections were recorded. Other variables included the teacher's estimates of success in each subject and the headmaster's estimate of success in the course of higher education selected by the student. Sociological data have been obtained by questionnaires.

The analysis of such a complicated mass of material has taken longer than was expected, but the first tables are now being produced by the computer. From them we hope to obtain answers to some questions, such as "How do students entering arts faculties differ from those who take up science or medicine or applied science?" The survey data have also provided a base line from which a follow-up project has been launched and which continues.

General Observations

From the preceding account it will be obvious that the Council has been fortunate in the cooperation it has received from schools, colleges, and universities. The expected response rate for its investigations is between 90 and 100 percent. In part this is because of the consultations among those concerned which precede the launching of a survey. In part it is because of the existence of a strong system of public schools. In part it is doubtless because of the size of the country, in which it is possible for many of these concerned in research, administration, and teaching to know each other personally.

The Council has also been fortunate in the financial assistance it has obtained. Bodies which have given substantial grants, without which these surveys could not have been conducted, include the Nuffield Foundation, the Eugenics Society, the Carnegie Trust for the Universities of Scotland, and the Scottish Education Department.

At the beginning of this paper there was a reference to the favorable climate towards educational research in the thirties. At times I have doubts whether the climate is so favorable now, at least so far as surveys are concerned. The advent of the comprehensive school, which most of us welcome, has brought with it in some quarters the attitude that tests and examinations are undesirable and that statistical analysis misses the whole point of education. A circular issued last year to education authorities by the Secretary of State for Scotland asked them to discontinue, wherever possible, external tests as measures of pupils' attainments at the end of primary schooling. The secondary school might, he considered, reasonably ask for an assessment of the pupils' attainments in the basic language and mathematical skills, but this was not to be done by using external tests, nor should the primary school devise internal examinations specifically for the purpose of preparing information about the pupil for the secondary school.

This hostility to external tests and examinations may be only a temporary phase. It is an interesting corollary that the number of students presenting themselves for the Scottish Certificate of Education, which is now awarded by an Examination Board on the results of external examinations taken by students towards the end of the secondary school, has increased by 75 percent in the last five years.

Whatever our interpretation of the situation, it is evident that anyone wishing to conduct an educational survey in Scotland today must, more than ever, be prepared to justify the proposal in terms of its purpose, the use to be made of its findings, and the load it will place on schools.

REFERENCES

These reports by the Scottish Council for Research in Education are given in order of publication. The publisher is the University of London Press Limited, London.

1. *The intelligence of Scottish children: a national survey of an age-group*, 1933.
2. Macmeeken, A. M. *The intelligence of a representative group of Scottish children*, 1939.
3. *The trend of Scottish intelligence*, 1949.
4. *Social implications of the 1947 Scottish mental survey*, 1953.
5. *Educational and other aspects of the 1947 Scottish mental survey*, 1958.
6. Macpherson, J. S. *Eleven-year-olds grow up*, 1958.
7. Maxwell, J. *The level and trend of national intelligence*, 1961.
8. *The Scottish scholastic survey, 1953*, 1963.
9. *The Scottish standardisation of the Wechsler Intelligence Scale for Children*, 1967.
10. *1953-63: Rising standards in Scottish primary schools* (on press).