To the Graduate Council:

I am submitting herewith a dissertation written by Canton Livingston Napier, entitled "A Computer-Based Test Analysis-Feedback System As A Supervisory Tool." I recommend that it be accepted in partial fulfillment of the requirements for the degree of Doctor of Education, with a major in Educational Administration and Supervision.

G.C. Ubben, Major Professor

We have read this dissertation
and recommend its acceptance:

John T. Lovell

John R Ray

Thomas C. Hood

Accepted for the Council:

L Evans Beck

Vice-Chancellor for
Graduate Studies and Research

Thesis
796
.N263
cop. 2

A COMPUTER-BASED TEST ANALYSIS-FEEDBACK SYSTEM AS A SUPERVISORY TOOL

A Dissertation

Presented for the

Doctor of Education

Degree

The University of Tennessee, Knoxville

Canton Livingston Napier

August 1979

1398076

Copyright by Canton Livingston Napier 1979
All Rights Reserved

ABSTRACT

The problem of this study was the need for more effective tools for a supervisory behavior system to utilize to influence the classroom testing behavior of teachers. The purpose of the study was to examine the effect of the implementation of a computer-based test analysis and feedback system on the testing behavior of teachers, as a first step in the development of a supervisory support system dealing with student evaluation.

Data as to test reliability, item difficulty, and item discrimination were collected from six tests for each of ten participants. Five participants that composed the experimental group received feedback from the system as to their test analysis, while the five members of the control group received no information. The participants were randomly assigned to groups.

The null hypotheses of no difference between means of the gain scores of the groups for test reliability, item difficulty, and item discrimination were not rejected at the .05 level of significance. Further, a time series analysis of the data required that no judgement be made as to the functional relationship of the dependent variables and time, as an independent variable.

In conclusion, analysis of the data did not support that feedback from the system directly to the individual concerned could influence the testing behavior of teachers.

TABLE OF CONTENTS

CHAPTER		PAGE
I.	PRESENTATION OF THE PROBLEM	1
	Introduction	1
	The Problem	3
	Introduction	4
	Statement of the problem	4
	Statement of the purpose	4
	The hypotheses	4
	Importance of the study	5
	Assumptions	11
	Definition of Terms	11
	Design of the Study	13
	The population	13
	The treatment	13
	Data gathering procedures	14
	Treatment and presentation of the data	15
	The design	16
	Test analysis-feedback system input	18
	Limitations or Delimitations	19
	Organization of the Study	19
II.	A REVIEW OF THE LITERATURE	20
	Introduction	20
	Change	20

CHAPTER PAGE

Test Item Analysis 34

Summary 43

III. THE DESIGN 46

Introduction 46

The Population 46

The Treatment 47

The Design 49

The Data Collection Technique 51

Presentation of the Data 52

Treatment of the Data 53

The Statistics 54

The Test Analysis-Feedback System 62

IV. PRESENTATION OF THE DATA 64

Introduction 64

Presentation and Analysis 65

Summary 83

V. SUMMARY, CONCLUSIONS, DISCUSSION, AND RECOMMENDATIONS 85

Summary 85

The Conclusions 87

Discussion 89

Recommendations 94

BIBLIOGRAPHY 96

CHAPTER PAGE

APPENDICES 102

Appendix A 103

Appendix B 113

Appendix C 115

Appendix D 120

Appendix E 128

Appendix F 151

VITA 153

LIST OF TABLES

TABLE		PAGE
I.	Observations of the Reliability Index by Subject by Test	66
II.	The Six Most Extreme Possible Outcomes in Either Direction for the Indices of Reliability	67
III.	Observations of the Difficulty Index by Subject by Test	69
IV.	The Six Most Extreme Possible Outcomes in Either Direction for the Indices of Difficulty	70
V.	Observations of the Discrimination Index by Subject by Test	72
VI.	The Six Most Extreme Possible Outcomes in Either Direction for the Indices of Discrimination.	73
VII.	Percent of Items Achieving Optimum Values for Item Difficulty	74
VIII.	Percent of Items Achieving Optimum Values for Item Discrimination	75
IX.	Population Parameters for Trend Analysis	79
X.	Population Parameters for Trend Analysis (Modified)	82

LIST OF FIGURES

FIGURE PAGE

1. Reliability Indices for Ten Subjects (E1 to C10). 129
2. Difficulty Indices for Ten Subjects (E1 to C10). . 131
3. Discrimination Indices for Ten Subjects (E1 to C10). 133
4. Reliability Indices with Optimum Value for Ten Subjects (E1 to C10) 135
5. Difficulty Indices with Optimum Value for Ten Subjects (E1 to C10) 137
6. Discrimination Indices with Optimum Value for Ten Subjects (E1 to C10) 139
7. Optimum Item Difficulty Achieved by Ten Subjects (E1 to C10). 141
8. Optimum Item Discrimination Achieved by Ten Subjects (E1 to C10) 143
9. Test Reliability for the Experimental Group . . . 145
10. Test Reliability for the Control Group 146
11. Item Difficulty for the Experimental Group 147
12. Item Difficulty for the Control Group 148
13. Item Discrimination for the Experimental Group . . 149
14. Item Discrimination for the Control Group 150

CHAPTER I

PRESENTATION OF THE PROBLEM

I. INTRODUCTION

Many public school systems maintain written philosophies and goals which imply direction for teachers within the classrooms. These written philosophies and goals are intended to influence the teaching behaviors of the staff, and to suggest the curriculum. Van Til, Vars, and Lounsbury stated that another force, the teacher constructed test, may more influence teaching and learning than does any other variable.

> After observing how examinations influence teaching and learning, some observers have concluded that "we teach what we test." By this they mean that testing is like the proverbial tail that wagged the dog. Instead of being one means of finding out how effective an educational program is, examinations sometimes prescribe the goals that will be sought, the content of the curriculum, and set limits on the methods and procedures used.[1]

Overly suggested that many students may possess a disdainful regard for teacher constructed tests, viewing them as having little value as a learning aid. He

---

[1]William Van Til and others, Modern Education for the Junior High School Years (New York: The Bobbs-Merrill Company, Inc., 1961), p. 470.

outlined an examination game that is played within the classroom by both the teacher and the student:

> Traditionally, too, assessment has been viewed as a game played by both the teacher and the student, with clearly defined roles for each. The role of the teacher has been to keep secret the questions she planned to ask on a test trying to trick the student into studying all of the assignments. The role of the student has been to attempt to outguess the teacher, studying only the material which he was sure would appear on a test.[2]

Kenney stated the importance of fostering a different emphasis toward tests and testing by suggesting criteria for an integrated information system which would support individualization of instruction.

> Teachers, in the usual environment of the classroom, could probably achieve a much higher degree of individualized instruction than they now achieve. The authors have known a few "master teachers" who were able to achieve this to a high degree. But in order to accomplish the kind of instruction we are talking about, an integrated information system is a necessary requirement. The teacher must have solid, reliable, relevant information. Information about the present status of a pupil, information about alternative instructional objectives, evaluative information regarding program effectiveness, and information providing possible clues to failures are all types of information that could prove useful to the teacher.[3]

If the effectiveness of an educational program is to be defined in terms of student success on teacher constructed

---

[2]Donald E. Overly and others, The Middle School: Humanizing Education for Youth (Worthington, Ohio: Charles A. Jones Publishing Company, 1972), p. 55.

[3]James B. Kenney and R. Robert Rentz, Automation and Control of Public School Instructional Records (Itasca, Illinois: F.E. Peacock Publishers, Inc., 1970), p. 97.

examinations, and other variables, then one may conclude that educators should attempt to insure the validity and reliability of their tests, qualities which may require more time and skill than many teachers possess.

II. THE PROBLEM

Introduction

In theory, the principal has been the instructional leader for a school, that is, the person most responsible for the effectiveness of the various parts of a total program. This role requires interaction with the members of the staff. Yet, increasing demands upon the time of principals as administrators tend to limit their ability to function within a supervisory behavior system.

While there are many processes, roles, or functions of supervision, within the present system of public education a competent building supervisor must be aware of the key function of communication. Hence, the need for data to communicate becomes apparent when a supervisor attempts to assist teachers to become more effective. And yet, very little activity within present systems of evaluation provides the process for a supervisory behavior system to obtain data about daily teaching behavior such as testing.

Statement of the Problem

The problem of this study then is the need for more effective tools for a supervisory behavior system to utilize to influence the classroom testing behavior of teachers.

Statement of the Purpose

The purpose of the study was to examine the effect of a Test Analysis-Feedback System on the testing behavior of teachers, as the first step in the development of a supervisory support system dealing with student evaluation.

The Hypotheses

To give direction to the study, the following hypotheses were formed:

(1) There will be no difference between the mean of pretreatment-posttreatment gain scores of the experimental group and the control group for the reliability indices.

(2) There will be no difference between the means of pretreatment-posttreatment gain scores of the experimental group and the control group for the difficulty indices.

(3) There will be no difference between the means of pretreatment-posttreatment gain scores of the experimental group and the control group for the discrimination indices.

The null hypotheses were rejected at the .05 level.

Importance of the Study

Bloom suggested the feedback "effect of examinations may help the teacher discover both strengths and weaknesses in the curriculum and in the learning experiences."[4] He wrote that it consists of item analysis data that is:

> made available to teachers in a form which they can relate to the learning experiences, materials, and content of instruction, and to sequence of learning experiences, the teachers are likely to make appropriate modifications in the curriculum and organization of instruction.[5]

Concluding, Bloom urged:

> Examinations are not an end in themselves. The examining process must be viewed as a means of making the educational process more effective.
>
> . . . The kinds of examinations used and the ways in which they are used and regarded by teachers and students (as well as administrators and parents) may have powerful effects on what is learned and how it is learned.[6]

As of yet, there are few techniques which realistically permit a supervisor to monitor and control teaching behavior such as classroom testing. Anderson suggested patterns of control of sub-systems within formal organizations such as public schools: direct supervision, impersonal mechanisms, professional standards, and bureaucratic rules. Although many complex organizations can rely

---

[4]Benjamin S. Bloom, "Testing Cognitive Ability and Achievement," Handbook of Research on Teaching, ed. N. L. Gage (Chicago: Rand McNally and Co., 1963), p. 393.

[5]Ibid., p. 394.

[6]Ibid., p. 395.

mainly on one for control of effectiveness, he suggested that all four patterns of control are to be found within the sub-systems of educational organizations.[7] The use of the concept of impersonal mechanisms may have utility for control if a school may be conceptualized as a social system.

A problem with such an assumption is reflected in the gap between the model of the school as a social system and many of the behaviors within a school. Although the dynamics of the Getzels and Thelen model showing the classroom as a social system suggest the need for constant feedback,[8] now, as in the past, members of the school organization, or its clients, may ignore totally, ignore in part, or accept information about behavior within the sub-systems. With new techniques and technology, it may be possible to identify evaluation procedures which will enable a supervisory behavior system to better understand, predict, and control teaching behavior.

The Association for Supervision and Curriculum Development, in their 1967 Yearbook, asserted that the proper role of evaluation was that as an agent for feedback and

---

[7]James G. Anderson, Bureaucracy in Education (Baltimore: The Johns Hopkins Press, 1968), pp. 1-39.

[8]Jacob W. Getzels and Herbert A. Thelen, "The Classroom Group as a Unique Social System," The Dynamics of Instructional Groups: Sociopsychological Aspects of Teaching and Learning, Fifty-ninth Yearbook of the National Society for the Study of Education, Part II (Chicago: University of Chicago Press, 1960), p. 80.

and guidance. The yearbook develops the theme that evaluation of our programs must be preceded by the elimination of our present form of record keeping and student evaluation.[9]

Mitzel, while discussing the concept of evaluation of teacher effectiveness, noted:

> More than a half century of research effort has not yielded meaningful, measurable criteria around which the majority of the nation's educators can rally. No standards exist which are commonly agreed upon as the criteria of teacher effectiveness.[10]

Lewis further cautioned:

> They (recent studies) show that there is a very little, if any, positive relation between the ratings of teachers and achievement of their pupils. In other words, a trained observer, using available instruments, cannot tell how well a teacher's pupils are learning by observing his performance in the classroom.[11]

Allen suggested that personality patterns are not useful for predicting success,[12] and Ryans emphasized,

---

[9]Association for Supervision and Curriculum Development, Evaluation as Feedback and Guide, 1967 Yearbook (Washington, D.C.: The Association's Press, 1967), p. 6.

[10]Harold E. Mitzel, "Teacher Effectiveness," Encyclopedia of Educational Research, ed. Chester W. Harris (3d ed.; New York: Macmillan Company, 1960), p. 1481.

[11]Gertrude M. Lewis, The Evaluation of Teaching (Washington, D.C.: National Education Association Publications, 1966), p. 66.

[12]Paul M. Allen and others, Teacher Self-Appraisal: A Way of Looking Over Your Own Shoulder (Worthington, Ohio: Charles A. Jones Publishing Company, 1970), p. 16.

concerning successful teaching, that little is yet known about adequate criteria of effectiveness.[13]

If direct supervision, professional standards, and bureaucratic rules offer such limited input into an evaluation system, then examination of Anderson's impersonal mechanisms may be considered a worthy project. As an impersonal mechanism, testing programs within schools, due to their ever felt presence, may have utility for program control. And yet, as suggested below, our testing programs may have definite limitations.

Feyereisen and others reflected that since the teacher education programs do not prepare teachers for their role in measurement and evaluation, the school systems will have to provide adequate in-service education if the staff is to effectively use examinations as tools to help attain the goals of the school system. Assuming poor measurement and evaluation skills, they stated:

> Until this situation is remedied school systems must assume the reponsibility (sic) for providing teachers with the knowledge and skills they need to participate effectively in the evaluation process.[14]

---

[13] David G. Ryans, "Problems in Identifying Effective Teaching," Growth, Teaching, and Learning: A Book of Readings, eds. H.H. Remmers and others (New York: Harper and Brothers, 1957), p. 539.

[14] Kathryn V. Feyereisen and others, Supervision and Curriculum Renewal: A Systems Approach (New York: Appleton-Century-Crofts, 1970), p. 316.

Schoer[15] agreed that test construction, although one of the most complex teaching tasks, receives little emphasis during teacher education. Once in service, he noted, the time available for the teacher to study to improve evaluation skills is limited.

Bloom and others proposed a system to improve evaluation that required four assumptions to be met:

> First, an inservice training program which enhances the relevant skills of teachers; second, cooperation so that no one individual will have to carry too heavy a load; third, released time to work on the projects; and finally, leadership to direct and coordinate the work.[16]

Continuing, they proposed the use of basic data-processing equipment as a labor-saving technique for the clerical tasks of scoring and statistical work.[17]

Observation of testing behavior of teachers at all grade levels of the public schools seemed to suggest that little manipulation of test data occurs once a score from a test has been recorded. A well designed, computer-based

---

[15]Lowell A. Schoer, Test Construction: A Programmed Guide (Boston: Allyn and Bacon, 1970), p. vii.

[16]Benjamin S. Bloom and others, Handbook on Formative and Summative Evaluation of Student Learning (New York: McGraw-Hill Book Company, 1971), p. 250.

[17]Ibid., p. 253.

test feedback system could produce statistics that would permit the teacher and the supervisor to monitor progress toward organizational goals by a student, a grade level, a department, a teacher, or an entire school.

Baker suggested that individual data might best be merged for the production of any reports. Apparently, Baker would be more comfortable with a system that would not reveal data which could be associated with an individual. Either way, at present, such input is usually missing from decision-making activities.[18]

On final selection from the literature supporting the importance of the study, appeared in The Kappan. Addressing the role of computers in education, Dagnon and Spuck stated:

> One of the keys to efficiency and quality improvements in every sector of our society has been augmentation of human effort by technology. There are strong reasons to believe that educators' efforts will also be augmented by technology. The next 10 years have the potential for enormous change in education; however, the success of these changes will depend on the quality of planning done by education leaders.[19]

---

[18]Frank B. Baker, "The Role of Statistics," Computer-Assisted Test Construction, ed: Gerald Lippey (Englewood) Cliffs, N.J.: Educational Technology Publications, 1974), pp. 148-165.

[19]Carol Dagnon and Dennis W. Spuck, "A Role for Computers in Individualizing Education--and It's Not Teaching," The Phi Delta Kappan, LVIII (February, 1977), p. 462.

III. ASSUMPTIONS

The following assumptions were accepted for the study:

(1) In our society, tests function as a fundamental, therefore important, tool of education.

(2) Item difficulty, item discrimination, and reliability scores are quantitative descriptors of the goodness of a test, or test item.

(3) The perceived relative advantage, that is, the cost-benefit or other advantages of the Test Analysis-Feedback System over existing practices will be such as to insure that sufficient time and other resources will be utilized to improve testing behavior.

(4) The magnitude of the difference of the gain scores has utility as an indicator of change.

(5) Tests will be constructed by the teacher after intervention of the feedback from the previous test.

(6) The quantity of information presented by the Test Analysis-Feedback System, and the format in which the information is presented, are optimums.

(7) The information presented by the system was understandable by the teachers.

IV. DEFINITION OF TERMS

Teacher. An individual within a school who has instructional, but not supervisory, duties.

Supervisor. An individual within a school who has supervisory, but not teaching, duties.

Feedback. Feedback is information from the results of a behavior that has the potential to influence the actions of an individual.

Test Analysis-Feedback System. A Test Analysis Feedback System is a technique for providing information that may be used to evaluate a test or test item. The design permits data manipulation with minimum user effort.

Test. A test is an instrument used to determine achievement of content, or skills. Alternatives presented as possible answers must be limited to four categories, usually A, B, C, D; 1, 2, 3, 4; or, true, false. Such items must compose two-thirds of the instrument.

Testing Behavior. Actions by teachers that influence test reliablity, item difficulty, and item discrimination, indices.

Item Ease. The percent of all examinees who answered an item correctly.

Item Difficulty. A quantity that expresses how many students who scored in the top, and bottom, 27 percent of a class incorrectly answered an item. Scores may range from 0 percent to 100 percent.

Item Discrimination. A quantity that compares how well high scoring individuals did on an item, as compared to individuals who scored poorly on the test, with the high and

low groups each containing 27 percent of the students. Discrimination for an item may be reported from minus to plus one (-1, to +1).

Reliability Index. A coefficient obtained from the Kuder-Richardson "formula 20" that reports the internal consistency of a test.

Item Difficulty Index. The mean for the item difficulty of all items on a test.

Item Discrimination Index. The mean, multiplied by 100, for the item discrimination of all items on a test.

V. DESIGN OF THE STUDY

The Population

The population was composed of ten teachers. To form a treatment and a control group, members of the population were randomly assigned to two samples, each with an N of five.

The Treatment

Procedures for implementation of the Test Analysis-Feedback System into classroom test administration were presented individually to the members of the experimental group in written form. Members of the experimental group were treated with the following information:

(1) A raw score of the number of correct responses for each student.

(2) The percent of correct responses for each student.

(3) An item analysis that showed the student response for each question as correct or incorrect.

(4) Information about the class average.

(5) A graph of item ease.

(6) The number of responses to each alternative for each item.

(7) An indicator of test reliability.*

(8) The item difficulty for each item.*

(9) The item discrimination for each item *

for six tests. (See Appendix A.)

Information marked, *, was accompanied by written suggestions as to how to attain the optimum score. (See Appendix C.) Except for the request to participate, members of the control group received no information. (See Appendix B.)

Data Gathering Procedures

From the experimental group, copies of all outputs that were generated by the system after each test analysis were retained for the study. Observations of the reliability index, the item difficulty index, and the item discrimination index, were made.

From the members of the control group, copies of the answers were obtained for six tests. Although the tests were analyzed by the system, no information was given to the members of the control group. Observations of the reliability index, the item difficulty index, and the item discrimination index, were made.

Treatment and Presentation of the Data

The randomization test for two independent samples was chosen to test the significance of the difference between the gains of the two independent samples. The test does not require that the assumptions of normal distribution or homogeneity of variance be made. The test does require at least interval measurement of the variables.

Siegel stated that the test is, "a useful and powerful nonparametric technique for testing the significance of the difference between the means of two independent samples when n^1 and n^2 are small."[20] He concluded his discussion of this test by stating, "Because it uses all the information in the samples, the randomization test for two independent samples has power-efficiency, in the sense defined, of 100 percent."[21]

---

[20]Sidney Siegel, Nonparametric Statistics for the Behavioral Sciences (New York: McGraw-Hill Book Company, Inc., 1956), p. 152.

[21]Ibid., p. 156.

Observations for analysis included the reliability index, the item difficulty index, and the item discrimination index. Test one was the pre-treatment observation selected for analysis, and test six was the post-treatment observation selected for analysis. The null hypothesis was rejected at the .05 level of significance.

Reliability, item difficulty, and item discrimination have an optimum value for the greatest contribution to the goodness of a test. (For example, on a scale of from 0 to 100, the optimum for item difficulty is 50.) Therefore, a descriptive analysis, in tabular and graphical form, as to degree and direction of change, was presented.

Further, it was possible to treat the observations as a time series. In effect, to determine the functional relationship of time and the observations, the study specified a task to be completed (a change in testing behavior), then examined the time required to complete the task through simple regression analysis.

The Design

Campbell and Stanley[22] stated that their Pretest-Posttest Control Group Design,

---

[22]Donald T. Campbell and Julian C. Stanley, "Experimental and Quasi-Experimental Designs for Research on Teaching," Handbook of Research on Teaching, ed. N.L. Gage (Chicago: Rand McNally and Co., 1963), pp. 183-194.

$$\begin{array}{cccc} R & O_1 & X & O_2 \\ R & O_3 & & O_4, \end{array}$$

controls for classes of extraneous variables that might effect the internal validity of an experiment. They listed these possible extraneous variables: History, maturation, testing, instrumentation, statistical regression, selection, and experimental mortality.

The above design controls for the effect of history as any event that might have caused a difference between O_1 and O_2 would also have caused a difference between O_3 and O_4. In this study, history was further controlled by the participants not being brought together in a single session where one unique event for either the experimental or the control group could become a possible explanation of the difference between the pre and post observations.

Maturation, processes internal to an individual, was controlled in that such processes would influence both groups, as would the effects of any testing. Observations were obtained by one individual from the output of the Test Analysis-Feedback System, therefore the effects of instrumentation were controlled.

Random assignment to treatment or control conditions ruled out regression or selection as an explanation of the difference between groups. The ability to identify all scores by subject permitted determination of the influence of mortality.

Test Analysis-Feedback System Input

From the experimental group, data to be analyzed were read directly from card decks that had been stacked by students. The students were required, (1) to show their choice of alternatives by manipulation of one prepunched unit record card for each answer, and (2) to place the cards in proper sequence within their deck.

From the control group, data for analysis were also obtained from card input, with the difference being that the deck was not prepared by the students of teachers in the control group.

A unit record card may be punched in such a manner that two columns may be used to read four different values, depending upon which side of the computer card is turned up, and which end is first into the reader. This permits the four longer edges of the card, in this case identified as A, B, C, or D, and 1, 2, 3, or 4, to represent different alternatives. True and false alternatives may be identified by marking one edge true and one edge false.

By proper orientation and placement of a card within a card deck, student answers may be fed directly into a computer system without the use of a keypunch operator or a mark sense reader.

VI. LIMITATIONS OR DELIMITATIONS

The study was limited to teachers in schools that were located within one hour driving time from the campus of The University of Tennessee, Knoxville. Participants were volunteers who were asked to contribute to the study. Technical demands limited use of the system to tests of 50 questions or less.

VII. ORGANIZATION OF THE STUDY

Chapter I contains the introduction to the study, the problem, the assumptions, the definition of terms, the design of the study, and the organization of the study. Chapter II is a review of the related literature. Chapter III is a presentation of the design, with an introduction, the population, the procedures, the treatment, the data, and treatment and presentation of the data. Chapter IV is a presentation of the data. Chapter V contains a summary, conclusions, discussion and recommendations.

CHAPTER II

A REVIEW OF THE LITERATURE

I. INTRODUCTION

This review of the literature was organized with the following components: Change theory, and item analysis principles. The selected literature on item analysis principles was included to support the specifics of the treatment for the experimental group. Models for understanding change within an individual, and within a social system, were presented, with the specific function of feedback as a tool to influence behavior being stressed.

II. CHANGE

There have been a number of models proposed for use in the study of change.[1] There is general agreement that, having defined a problem, one may note there are many

---

[1]Kurt Lewin, Field Theory in Social Science (New York: Harper, 1951); see also Warren G. Bennis, Kenneth D. Benne, and Robert Chin (eds.), The Planning of Change (New York: Holt, Rinehart and Winston, 1962); A.H. Maslow, "A Theory of Human Motivation," Psychological Review, L (July, 1943), 370-396; Jacob W. Getzels and Herbert A. Thelen, "The Classroom Group as a Unique Social System," The Dynamics of Instructional Groups: Socio-Psychological Aspects of Teaching and Learning, Fifty-ninth Yearbook of the National Society for the Study of Education, Part II (Chicago: University of Chicago Press, 1960), p. 80.

avenues to a possible solution, or change. Solutions imply change in behavior. And, a change in behavior must be associated with changing people, not with changing things. Although learning a new skill, a thing, may be relatively easy, the desirable change in the behavior of the individual may require greater effort.

The existence of such models does not suggest that change is a desirable condition, but does imply that change exists. If we conceptualize supervisory behavior within a social system as operating to change behavior, we may explore the influence of factors operating to increase, or decrease, forces upon the individual.

A system may be explained, in a most rudimentary sense, in terms of inputs, throughputs, and outputs.[2] Adding to this simplicity, authors have suggested that other concepts are necessary to fully develop the concept of a system: Openness, boundary, external feedback, internal feedback, and sub-systems, to name a few. The reader may examine Bennis,[3] or Margulies,[4] for a more in depth conceptualization of a social system.

---

[2]Newton Margulies, and Anthony P. Raia, Conceptual Foundations of Organizational Development (New York: McGraw-Hill Book Company, 1978), pp. 3-54, 82-101.

[3]Warren G. Bennis, Kenneth D. Benne, and Robert Chin (eds.), The Planning of Change (New York: Holt, Rinehart and Winston, 1962), pp. 187-478.

[4]Margulies, Ibid., pp. 3-54, 82-101.

In the perfect world of theory, the concept of feedback insures that a system's output is directed efficiently and effectively toward a goal. Unfortunately, or fortunately depending upon your point of view, the real world may not permit perfect control of people as they function within an organization.

To accept this view, then conclude that exploration of a social system will be flawed beyond usefulness, is to suggest that change within an organization cannot be directed. Until evidence is presented to support that members of an organization do not behave as suggested by social systems theory, and with the awareness of the presence of complex interactions that may, or may not, be identifiable, the literature supports that research should be undertaken to understand the functions within a social system.

Parsons and Shils proposed that

> a social system is a system of the actions of individuals, the principal units of which are roles and constellations of roles. It is a system of differentiated actions, organized into a system of differentiated roles.[5]

It is possible to view feedback as one of those differentiated actions. To be useful as a tool for changing

---

[5]Talcott Parsons and E.A. Shills, Toward a General Theory of Action (Cambridge, Mass.: Harvard University Press, 1971), p. 197.

an individual's behavior toward organizational goals, feedback needs to be fully understood in terms of who, what, when, where, why, and how. Nadler concluded that

> there has been relatively little attention given to the "why" and "how" of data-based methods. As a result, little progress has been made in the development of an applied theory of information (in general) and feedback (in particular). At the same time, an examination of the work of researcher in social and organizational psychology reveals a rather large (if diffuse) body of knowledge about how information affects behavior. This research implies that the way in which information is collected and the manner in which it is fed back can have predictable effects on behavior.[6]

Nadler[7] stated that feedback has two functions that can effect behavior within an organization. It can energize behavior, that is, motivate the individual, the group, or the organization ot seek change. Such information can also direct behavior by lessening the alternatives for solving a problem.

Margulies and Raia referred to this as providing "impetus for problem solving," and providing an understanding of "alternatives for potential change."[8] This second

---

[6]David A. Nadler, Feedback and Organization Development: Using Data-Based Methods (Reading, Mass.: Addison-Wesley Publishing Company, 1977), p. 57.

[7]Ibid., p. 58.

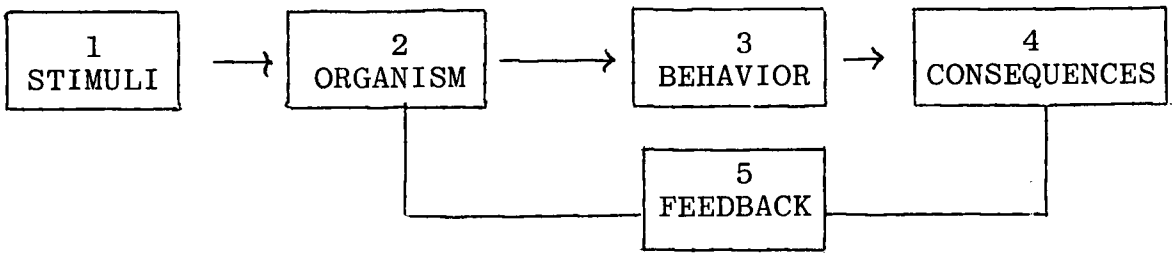
[8]Margulies, op. cit., p. 141.

function of feedback is described by Annett as the reduction of uncertainty, and is considered as central to a useful understanding as to how behavior may be modified, or how learning takes place.[9]

Margulies and Raia summarized a number of definitions of learning, then concluded:

> It is our belief that human learning also includes becoming aware of and experiencing the consequences of one's own behavior. This involves the notion of "feedback," which is the connecting link between learning and systems theory.[10]

Their model of the major components of learning suggests the importance of feedback to individuals or social systems. It was displayed as:[11]



Implications for this study from the above model were seen as questions of what consequences were to be presented, and when.

The reader who desires a comparison and contrast of some of the major learning theories may turn to Margulies

---

[9]John Annett, Feedback and Human Behavior (Baltimore: Penguin Books Inc., 1969), p. 139.

[10]Margulies, op. cit., pp. 83-84.

[11]Ibid., p. 84.

and Raia for their conclusions. They expanded their concern about learning from the individual to organizational learning, then suggested considering the organization as a social learning system which is characterized by "actual or implied presence of 'others' while learning occurs and to their participation in the process."[12]

Obviously aware of the controversy found in the literature about learning, Margulies and Raia formulated seven generalizations about learning. Of their seven, two gave direction to this study. They were:

> Fourth, desired behavior must be reinforced if it is to be learned. Reinforcement may occur in many ways and at different levels and includes a sense of progress toward desired goals. Fifth, feedback is an essential condition of learning. This enables the individual to assess the consequences of behavior.[13]

Stating that while the literature is exhaustive as to the nature and quality of feedback as it influences learning, and that there are numerous theories about the individual as the learner, Margulies and Raia wrote: "Only a small amount of the literature is concerned with the subject of social learning and even less deals with the question of organizational learning."[14]

---

[12] Ibid., p. 89.

[13] Ibid.

[14] Ibid., p. 84.

They proposed a model of an organization as an open sociotechnical system. To facilitate learning, their model conceptualized two sources of feedback for an organization, internal and external, with either having potential as a force for change.[15]

Kolb and Frohman[16] identified a model that permits an organization to learn and to improve during a typical change process. One stage of their model requires that previously collected data be used to motivate individuals or groups to unfreeze, to begin to change.

Margulies and Raia declared that "the three-phase model of unfreezing, moving, and refreezing has been the definitive tool of both theoreticians and practitioners involved in organizational change."[17] This model, by Lewin,[18] proposes that behavior can be explained as a balance between opposing forces; vary the strength of any one force to the proper degree, and the behavior will change. It is conceivable that by applying the model within a social system concept, one may propose alterations that have potential for directing the system toward a goal.

---

[15] Ibid., pp. 9-14.

[16] D.A. Kolb and A.L. Frohman, "An Organization Development Approach to Consulting," Sloan Management Review, XII (1970), 51-65.

[17] Margulies, op. cit., p. 132.

[18] Lewin, loc. cit.

It would appear that during any of the three stages feedback would be necessary. Without such information it is difficult to see how one could determine the degree or direction of change.

The use of feedback in different ways to facilitate achievement of organizational goals was documented by Nadler.[19] As examples, he suggested the collection and analysis of feedback by the working group that would use it, feedback to people who will not work together as in a T-group, and feedback to a subordinate group previous to a supervisor viewing it. Presenting these and others, he concluded that "other approaches exist and many more await to be developed and articulated."[20]

Heller proposed that within an organization the true effect of feedback may be better examined through his Group Feed-Back Analysis method. He found similarities between his method and others, but suggested his step of having equals examine feedback, that is, having no supervisors with a group, presents "new information in psychologically acceptable form, thereby facilitating the adoption of alternative points of view."[21]

---

[19]Nadler, op. cit., pp. 158-165.

[20]Ibid., p. 164.

[21]Frank A. Heller, "Group Feed-Back Analysis as a Change Agent," Human Relations, XXIII, 4, (August, 1970), 319.

Alderfer and Holbrook, having seen the same psychological limitations where the superior is present for initial discussion, proposed their alternative design for the family group feedback design that incorporates the superior and his immediate subordinates. Their peer group-intergroup model was conceived as having two phases. During the first phase, equals met to receive feedback, that is, supervisors with supervisors, and workers with workers. Only after each group interacted with the information was there formal interaction between members of the organization at different levels of authority.

They did not suggest that their model was better than other models, but did propose that "from a theoretical position that there are conditions when the peer group-intergroup design may be preferable to the family group design."[22] If so, one might conclude that present technological advances could create conditions where other models could be considered. An implication from the above is that one such model could propose no intervention by a supervisor at any stage of feedback analysis.

A study by Klein, Kraut, and Wolfson examined the impact on the attitude of recipients of feedback under varying conditions of structure and process. Aspects of

---

[22]Clayton P. Alderfer and John Holbrook, "A New Design for Survey Feedback," Education and Urban Society, IV (1972/1973), 463.

structure were seen as those activities that facilitated the individual being brought into contact with the feedback, while the conditions of process were the internal dynamics of the meeting. They concluded that "structure facilitates process and is its natural antecedent,"[23] while their "process variables were more powerful predictors of the dependent measures"[24] of attitude. Their study had implications for this project in that it suggested that the way a person can interact with information has more influence on his behavior than do the conditions under which it is presented.

French and Bell suggested, among other conditions, that feedback at the interpersonal level is more constructively received when it is specific, frequent, and delivered directly to the individual who can take whatever action is necessary.[25]

---

[23]Stuart M. Klein, Allen I. Kraut, and Alan Wolfson, "Employee Reactions to Attitude Survey Feedback: A Study of the Impact of Structure and Process," Administrative Science Quarterly, XVI, 4 (December, 1971), 497.

[24]Ibid.

[25]Wendell L. French and Cecil H. Bell, Jr., Organizational Development: Behavioral Science Interventions for Organization Improvement (Englewood Cliffs, N.J.: Prentice-Hall, Inc., 1973), pp. 161-162.

Collins and Guetzkow[26] asserted that within an organization there are two sources of feedback influencing behavior. They referred to these sources as interpersonal, or feedback from fellow organization members, and task-environment, or feedback from the environment as to goal achievement. Task-environment rewards were seen as having greater potential than interpersonal rewards, but were considered to be handicapped in many organizations by the amount of time it takes to obtain feedback about our role success.

An implication of their conclusion is that, since interpersonal feedback tends to be short range, while task feedback may in some cases take years, it is possible to envision organizational goals being directed more by what other members of a group feel or think, than by how well the task is being achieved for the organization. Apparently, techniques need to be developed that will provide more immediate feedback, directly from the environment to the individual.

For teachers, and supervisors, the above conceptualization suggested a more acute problem than that found within

---

[26]Barry E. Collins and Harold Guetzkow, A Social Psychology of Group Processes for Decision-Making (New York: John Wiley & Sons, 1964), pp. 85-87.

many other fields. There has tended to be little of either type feedback, therefore teachers have received little information about the success, or failure, of their pupils. Usually, individuals have not been consistently present to give either interpersonal, or task, feedback.

Discussing his experiment where individuals working on a group task received feedback as to the group's performance, their own performance, and the performance of other team members, in varying cycles, Zajonc concluded that, among other things,

> the most pronounced improvement occurs for a difficult group task when information about the performance of all team members as well as of the team as a whole is made available.[27]

that is, feedback about performance of a group had little effect on the performance of the individuals within the group unless information was also made available about each member. Apparently, status reports about department or school achievement would motivate less change within the organization than a process that in addition reported individual contributions.

"Effective and permanent adult learning comes after the individual has experimented with new approaches and

---

[27]Robert B. Zajonc, "The Effects of Feedback and Probability of Group Success on Individual and Group Performance," Human Relations, XV, 2 (May, 1962), 160.

received appropriate feedback in the on-the-job situation,"[28] confirmed Beer and Huse when they presented a number of findings from an experiment on a systems approach to organizational development.

Nadler, Mirvis, and Cammann provided some conclusions from their research with an ongoing feedback system that were relevant to the focus of this study. Noting that present systems of feedback within many organizations were often inadequate, they set forth to examine the possibility that better designed feedback systems could increase the effectiveness of an organization. Having indicated that a strong system should provide feedback about the technical and human factors within an organization, should provide relevant information, should be timely and addressed to the people who can use it, should be seen by managers as a tool for change rather than as a basis of personal evaluation, and should involve the user in the design or change, they presented what they indicated represented "a first step in the development of ongoing feedback systems as useful managerial tools."[29]

---

[28]Michael Beer, and Edgar F. Huse, "A Systems Approach to Organization Development," The Journal of Applied Behavioral Science, VIII, 1 (1972), 99.

[29]David Nadler, Philip Mirvis, and Cortlandt Cammann, "The Ongoing Feedback System: Experimenting With a New Managerial Tool," Organizational Dynamics, IV, 4 (Spring, 1976), 80.

Some implications from their experiment were accepted as important for this study. Specifically, they declared that

> feedback by itself can have positive effects if the employees who are to use the feedback have the skills necessary to interpret, understand, and use the data and if they are motivated to use the information on their own.[30]

But, they also found that in many organizations

> feedback will have a positive effect only when managers motivate employees to work with the information and assist them in using the data through meetings, through participation, and through the use of goal-setting techniques.[31]

"In most cases," Nadler, Mirvis, and Cammann found, "feedback by itself will not be sufficient to improve organizational performance."[32]

Nadler wrote an excellent conclusion for this section.

> The idea that feedback interventions can enhance organizational functioning is an appealing one. Feedback can be observed in both physical and social systems, and it would seem that under the right conditions feedback could be a potentially powerful tool in the improvement and transformation of organizations. The research and theory to date have given some indication of the potential of feedback interventions, but the surface has only been scratched. Examining the work to date and identifying new directions should form the basis of

---

[30] Ibid., p. 78.

[31] Ibid.

[32] Ibid.

more systematic work on developing data feedback as an effective change technology.[33]

III. TEST ITEM ANALYSIS

What phenomena of testing behavior may need improvement? Ebel[34] identified a number by suggesting that teachers make the following common mistakes.

They pay little attention to the reliability of their tests, relying a great deal on their subjective feelings. Teachers set absolute standards as to what is a passing and failing score on a test, not accepting that any one test could contain poor items, which in turn compose an instrument that is not quite that accurate. Test preparation is usually a hasty after thought to the instructional process. Tests do not sample all that was taught, usually due to their short length. Individual items test trivial facts. Components of test items have technical weaknesses. Little consideration is given to what the true score of an individual might be if sampling errors are considered. And, concluding, Ebel noted that tecniques of statistical analysis are seldom used to investigate the quality of tests.

---

[33]David A. Nadler, "The Use of Feedback for Organizational Change: Promises and Pitfalls," Group and Organization Studies, I, 2 (June, 1976), 184.

[34]Robert L. Ebel, Essentials of Educational Measurement (Englewood Cliffs, N.J.: Prentice-Hall, Inc., 1972), pp. 92-93.

To permit examination of the effect of the implementation of the Test Analysis-Feedback System on the testing behavior of teachers, three of the above factors were incorporated as the dependent variables. This section of the review suggested acceptable criteria, and techniques, for determining item difficulty, item discrimination, and test reliability. The theory on which each rests was not discussed as it was assumed for this study that item difficulty, item discrimination, and test reliability scores are quantitative descriptors of the quality of a test.

Ebel, proposing a criterion for quality, stated that

> reliability is often the most significant statistical measure of the quality of a classroom test. While validity is generally considered to be more important than reliability, the statistical validity of a classroom test cannot ordinarily be determined.[35]

Although recognizing that the validity of many classroom tests can not easily be determined, Ebel asserted that "if the items are highly relevant and well balanced and if the scores are highly reliable, the test will necessarily be highly valid as a measure of achievement."[36]

Among the possibilities for determining the reliability of test scores are at least four well-known methods for obtaining an estimation: Kuder-Richardson, test-

---

[35]Ebel, Ibid., p. 379.

[36]Ibid.

retest, equivalent forms, or split halves. Considering the limitations within individual classroom testing program, the Kuder-Richardson "formula 20" was most appropriate for this project.[37]

Robert Thorndike, in an excellent presentation on reliability, set forth that "the fact that repeated sets of measurements never exactly duplicate one another is what is meant by 'unreliability,'"[38] while repeated measurements of the same object will tend to be somewhat consistent, "the reverse of the fact of variation which we have just considered, and will be designated 'reliability.'"[39] As tests are instruments of measurement

> the lower the reliability of the score, the more tentative the judgment or decision must be, until in the extreme case, as the reliability approaches zero, the score provides absolutely no basis at all for any judgment or decision.[40]

With the caution that validity must never be sacrificed to obtain reliability, Thorndike concluded that data about

---

[37] G.G. Kuder and M.W. Richardson, "The Theory of the Estimation of Test Reliability," Psychometrika, II (September, 1937), 158.

[38] Robert L. Thorndike, "Reliability," Educational Measurement, ed. E.G. Lindquist (Washington, D.C.: American Council on Education, 1951), p. 560.

[39] Ibid.

[40] Ibid., p. 562.

reliability must be available as an aid in the interpretation of results.[41]

Ebel,[42] and the Educational Testing Service,[43] found that test publishers have traditionally considered good test reliability to be 0.90 or better. Ebel considered a somewhat lower reliability of 0.80 to be acceptable for classroom tests, while the writer for the Educational Testing Service accepted reliabilities between 0.60 and 0.80 as good.

If the objective of a test is to differentiate between levels of achievement within a group, then the test maker needs information as to which sample of items have the most potential for contributing to the improvement of the reliability of a test. Davis suggested that such information as item difficulty and item discrimination are effective tools for such selection. He summarized his position as

> construction of reliable or valid tests that are also efficient measuring instruments requires use of quantitative data concerning the difficulty

---

[41] Ibid., p. 619.

[42] Ebel, op. cit., p. 378.

[43] Educational Testing Service, Short-Cut Statistics for Teacher-Made Tests, Evaluation and Advisory Series (Princeton: Educational Testing Service, 1960), pp. 31-32.

> and discriminating power of each one of the pool of items available.[44]

He had previously indicated that for an item to maximally discriminate between examinees, the item should have a difficulty of 50 percent, as this level will permit the maximum number of discriminations to be made. Therefore, while it may be impossible to produce a test where all items are of optimum difficulty, to assure as much discrimination as possible among examinees, he wrote that the test should be constructed of items clustered around the 50 percent difficulty level.[45]

Using theoretical assumptions as a basis, Gulliksen demonstrated the usefulness of the item difficulty concept, and suggested an optimum difficulty level for an item.

> In order to maximize the reliability and variance of a test, the items should have high intercorrelations, all items should be of the same difficulty level, and the level should be as near 50% as possible.[46]

---

[44]Frederick B. Davis, "Item Analysis in Relation to Educational and Psychological Testing," Psychological Bulletin, XLIX, 2 (March, 1952), 117.

[45]Frederick B. Davis, "Item Selection Techniques," Educational Measurement, ed. E.G. Lindquist (Washington, D.C.: American Council on Education, 1951), pp. 308-311.

[46]Harold Gulliksen, "The Relation of Item Difficulty and Inter-Item Correlation to Test Variance and Reliability," Psychometrika, X, 2 (June, 1945), 79.

The ability of an item to discriminate between students who have met the objectives of a course and those who have not reflects that item's discrimination ability. The single most important factor influencing discrimination potential of an item is the difficulty level of the item according to Schoer.[47] He required that the item should be neither too easy, nor too difficult, as in neither case would the item function to discriminate.

To add importance to the proper selection of items, Ebel declared that

> the selection of highly discriminating items, using total test scores as the criterion, results in a test whose items are valid measures of what the whole test measures. In this sense, item analysis is a technique of item validation.[48]

The criteria for judging the quality of an item discrimination score was reported, but not justified, by Ebel.[49] He suggested the following quality for the range of scores given: High as .40, and up; moderate as .20 to .39; low as .01 to .19; and, zero, or negative, as not discriminating.

One technique for determining item difficulty, and item discrimination, requires using the scores obtained by

---

[47]Lowell A. Schoer, Test Construction: A Programmed Guide (Boston: Allyn and Bacon, Inc., 1970), p. 147.

[48]Ebel, op. cit., pp. 392-393.

[49]Ibid., p. 361.

the examinees as an internal criterion. Kelley[50] proposed that by selecting the examination papers of 27 percent of the lowest scoring students, and then selecting 27 percent of the highest scoring students, one is assured the greatest opportunity of making the groups as different as possible, while making the groups as large as possible.

Remmers and Gage, laying the foundation for evaluation of standardized and teacher-made tests, reported validity as a basic consideration when selecting or creating a test. A test was seen as valid if it met the criteria of relevance and reliability. The criterion of reliability, they noted, can be determined by computing coefficients from parallel forms of a test, from administering a test for a second time to the same group, or from examining one test for internal consistency.[51] To Remmers and Gage, the component of relevance was manifested from theory, curriculum analysis, relation to other tests, and the ability of the test to predict success.

---

[50]Truman L. Kelley, "The Selection of Upper and Lower Groups for the Validation of Test Items," Journal of Educational Psychology, XXX, 1 (January, 1939), 17-24.

[51]H.H. Remmers and N.L. Gage, Educational Measurement and Evaluation (rev. ed.; New York: Harper & Brothers, 1955), pp. 121-124.

Their characterization requires that a test be relevant and reliable before it can be considered valid. As it is considerably easier to determine a coefficient for reliability of a teacher-made test than it is to lay the groundwork for validity, one may argue that a first step in determining the quality of a test is to seek its reliability for that group.

Remmers and Gage did not choose to be precise when stating the value the reliability coefficient should reach. They commented that standardized test publishers usually require at least .80 in the population for which a test is designed, and found it sufficient just to conclude that the logic of the use of reliability requires that test makers seek to attain plus 1.0.[52]

A chapter on validity by Cureton[53] suggested that within the bounds of the resources of the individual teacher, at the present time, most attempts at determining validity, or how well the test serves its stated purpose, will have to be determined by the eyes of the beholder. Although the system used to analyze tests for this study could contribute

---

[52] Ibid., p. 141.

[53] Edward E. Cureton, "Validity," Educational Measurement, ed. E.G. Lindquist (Washington, D.C.: American Council on Education, 1951), pp. 621-694.

reliability information, such input would have to be used in conjunction with the concept of relevance to obtain an estimation of the validity. But, if we may assume that teachers would not knowingly administer a test composed of irrelevant items, then reliability scores may function to permit judgments as to the quality of a test, or testing program.

Although criterion-referenced tests and norm-referenced tests report achievement in different terms, Ebel characterized both as needing item discrimination information to help determine item quality as

> even the items in a criterion-referenced or mastery test are intended to discriminate those who have reached the criterion, or have achieved mastery, from those who have not.[54]

Although the previous statement accepts the difficulty of using traditional norm-referenced analysis procedures to examine criterion-referenced tests, it supports procedures for getting at the quality of a single test item, regardless of the stated approach of the teacher.

---

[54]Robert L. Ebel, "Can Teachers Write Good True-False Test Items?," Journal of Educational Measurement, XII, 1 (Spring, 1975), 35.

IV. SUMMARY

The testing behavior of teachers reflect a number of common mistakes as reported by Ebel. There are a number of observations and statistical techniques available that will permit examination of testing behavior to determine if the mistakes are distracting from the quality of the test.

Reliability was described as a criterion of the quality of a test by Ebel and Thorndike. The Educational Testing Service and Ebel determined acceptable reliability for a classroom test at less than that traditionally required by test publishers.

Item difficulty, and item discrimination, were reported as two factors that contribute to test reliability, each being seen as having an optimum value. A technique, from Kelley, for selection of groups to facilitate getting item difficulty and item discrimination was reported.

Determination of validity for classroom tests was reported as mostly resting on obtaining an acceptable reliability score, which was to be followed by examination of the items for relevancy.

From the wealth of literature about change and feedback, literature was reported about the function of feedback within a social system. The literature was found to contain many reports with opinions about models that should be

applied to bring change within an organization, yet quantitative data to support the usefulness of such models was difficult to find.

Specifically, the literature supported that the way feedback is collected within an organization, and then fed back, can have predictable effects on behavior.

Nadler, Margulies, and Raia reported that feedback can motivate and direct an individual, while Annett noted only that it can direct an individual.

Margulies and Raia reported that learning must be reinforced, and that feedback is an essential condition of learning. They expressed a need for research into social and organization learning.

As a technique, Heller, Alderfer, and Holbrook proposed that only equals should be present during the initial discussion of feedback.

Klein, Kraut, and Wolfson determined that how feedback may be used is important to organization members. French and Bell supported that feedback needs to be specific, frequent, and delivered to the person who can use it.

Collins and Guetzkow found time to be a factor controlling the power of feedback.

Nadler, Mirvis, and Cammann suggested that ongoing feedback systems could increase the effectiveness of an

organization. Further, in some cases, feedback directly to the concerned individuals, with no intervention by a supervisor was seen as a successful technique for improvement.

Zajonc, Beer, and Huse found feedback directly to the concerned individual to be effective.

The literature stated that item analysis information may be used to report the quality of a test. Such information, if conditions within an organization permit change to be based on it, may be considered as feedback.

Within a social system, such feedback directly to the concerned individual has the potential to bring change with or without supervisory, or other, intervention.

The degree of change is expected to be influenced by the characteristics of the data, the method of presentation, and the ability of the individual to react to it.

CHAPTER III

THE DESIGN

I. INTRODUCTION

This chapter contains the following sections: A description of the population with the method of selection, the treatment, the design, the data collection technique, how the data were presented, the methods of analysis, an introductory description to the statistics applied, and a description of the computer-based Test Analysis-Feedback System.

II. THE POPULATION

The population was composed of ten teachers. Although personal characteristics were not a consideration for membership in the population, the make-up of the population is presented to permit the reader to get a feel for the project.

The following school systems had participants in the project: Anderson County, Clinton City, Knox County, and Oak Ridge, all in Tennessee. There were seven males, and three females. All participants had been teaching three or more years. Although some participants taught

other subjects, six volunteered to submit social studies tests, while four submitted science tests. The number of students enrolled ranged from 18 to 79; with one fifth; four sixth; one seventh; one eighth, ninth, tenth, and eleventh grade levels.

After obtaining permission to implement the study from the appropriate individual in the above school systems, schools were selected. Criteria for selection of schools within these systems were relative closeness in driving time to the main campus of The University of Tennessee at Knoxville, and more importantly, identification of one teacher within each school who had a propensity toward administering tests containing items of the type that could be analyzed by the Test-Analysis Feedback System.

The identified teachers were asked to volunteer for a project concerned with the development of a flexible and inexpensive system that would provide teachers with information about classroom testing. All contacted individuals agreed to participate.

III. THE TREATMENT

After random assignment to treatment and control groups, the teachers received the appropriate written directions. The experimental group received the information

contained in Appendix A, while the control group received the information contained in Appendix B.

Appendix A contains the procedures for implementation of the Test Analysis-Feedback System into the classroom. In the form of a booklet, these instructions were presented individually to the members of the experimental group. Arrangements were made for picking up and returning the material described in the booklet. The output was returned by the next school day.

The following output was presented on what is referred to as "green bar" printer paper. No other information, such as technical information about the program, computer center announcements, and so on, was made available to the participants.

Each section contained the name of the school, the teacher, the subject, and appropriate headings for each of the following:

(1) A raw score of the number of correct responses for each student.

(2) The percent of correct responses for each student.

(3) An item analysis that showed the student response for each question as correct or incorrect.

(4) Information about the class average.

(5) A graph of item ease.

(6) The number of responses to each alternative for each item.

(7) An indicator of test reliability.*

(8) The item difficulty for each item.*

(9) The item discrimination for each item.*

Information marked, *, was accompanied by written suggestions as to how to attain the optimum score. Feedback was presented to members of the experimental group after each administration of five tests to facilitate the application of the randomization test for two independent samples.

IV. THE DESIGN

To form a treatment and a control group, members of the population were randomly assigned to two samples, each with an N of five. For purposes of random assignment, each member of the population was assigned a number from one to ten. A table of random numbers was entered to assign the individuals to the groups.

Campbell and Stanley[1] stated that their Pretest-Posttest Control Group Design,

---

[1]Donald T. Campbell and Julian C. Stanley, "Experimental and Quasi-Experimental Designs for Research on Teaching," Handbook of Research on Teaching, ed. N.L. Gage (Chicago: Rand McNally and Co., 1963), pp. 183-194.

R	O_1	X	O_2
R	O_3		O_4

controls for classes of extraneous variables that might effect the internal validity of an experiment. They listed these possible extraneous variables: History, maturation, testing, instrumentation, statistical regression, selection, and experimental mortality.

The above design controls for the effect of history as any event that might have caused a difference between O_1 and O_2 would also have caused a difference between O_3 and O_4. In this study, history was further controlled by the participants not being brought together in a single session where one unique event for either the experimental or the control group could become a possible explanation of the difference between the pre and post observations.

Further, after being identified as a possible participant, each teacher was individually, and separately, contacted. At no time during the study were they made aware of the identity of other participants. Central office personnel who were aware of the location of the various schools were asked not to locate or communicate with the participants concerning the study. The request was made to insure that no non-random, or unique, event for either the experimental or control group would occur as the result of such interaction.

Maturation, processes internal to an individual, was controlled in that such processes would influence both groups. Although this design would control for the effects of any testing, the rival hypothesis of testing was absolutely controlled as there was no testing of either group.

Observations were obtained by one individual from the output of the Test Analysis-Feedback System, and no changes were made in the measuring instrument, or observation technique, therefore Campbell's effects of instrumentation were controlled.

Random assignment to treatment or control groups ruled out statistical regression, or selection, as an explanation of the difference between the groups. The ability to identify all scores by subject permitted determination of the influence of mortality.

V. THE DATA COLLECTION TECHNIQUE

From the experimental group, copies of all outputs that were generated by the system were retained for the study. Observations of the reliability index, the item difficulty index, and the item discrimination index were obtained from the system output. From both groups, the first test provided pre-treatment data, and the sixth test provided post treatment data. Data from outputs one through six were retained for the time series analysis.

Members of the control group submitted the student answers, along with the teacher master, for direct examination. After the input for the system was obtained, the material was returned to the teacher.

VI. PRESENTATION OF THE DATA

The first step in presentation of the data was to form tables and graphs. The following tables were prepared to help with the determination of any possible relationships between the variables.

Tables were presented that compared the observed scores of the Reliability Index, the Item Difficulty Index, and the Item Discrimination Index, from six test analyses for each of the ten members of the population.

Graphs were prepared for each of the above observations. A graphical presentation permitted visual inspection for any unique event. Any observation that did not regularly change would become highly visable on the graph.

A second series of tables contained the pre-treatment and post-treatment observations of the Reliability Index, the Item Difficulty Index, and the Item Discrimination Index for the control group and the experimental group. Those observations that required rejection of the null hypothesis were identified in these tables.

Reliability, item difficulty, and item discrimination have an optimum value for the greatest contribution to the quality of a test. Graphs were presented that contained the observed scores, as in Figures 1, 2, and 3, but with the optimum values also identified. These figures, (4, 5, and 6), permitted visual inspection as to degree and direction of change of each index by each participant. (See Appendix E.)

Figures 7 and 8 reported the percentage of items that were in the desired difficulty and discrimination range for each test for each participant, Figures 4 through 8 were included to suggest quality changes that might not have been observable in the previous table displays. (See Appendix E.)

Further, tables and graphs were presented to permit visual examination of the equation, (slope, and intercept values), that suggested the best fitting line when the data were treated as a time series. Presentation of the data as a scatter diagram formed a pictorial that suggested an appropriate regression model for the time series.

VII. TREATMENT OF THE DATA

The randomization test for two independent samples was chosen to test the significance of the difference between the gains of the two independent samples. The test does not require that the assumptions of normal distribution,

or homogeneity of variance, be made. The test does require at least interval measurement of the variables.

Observations for analysis included the Reliability Index, the Item Difficulty Index, and the Item Discrimination Index. Test One was the pre-treatment observation that was selected for analysis, and test Six was the post-treatment observation that was selected for analysis. The null hypothesis was rejected at the .05 level of significance.

Further, it was possible to treat the observations as a time series. In effect, to determine the functional relationship of time and the observations, the study specified a task to be completed (a change in testing behavior), then examined, through simple regression analysis, the time required to complete the task.

VIII. THE STATISTICS

The following introductory comments to the statistics of the study were included for one who may wish a review of the randomization test for two independent samples, and simple regression analysis as it may be applied to a time series. The discussion of the randomization test relied

heavily on Moses, and Siegel,[2] while Dickinson,[3] Mills,[4] and Neter[5] were understandable in their presentation of simple regression analysis with its utility for time series analysis.

The design of this project required the examination of data from two independent samples. Each sample was composed of an N of 5. The null hypothesis was rejected at the .05 level of significance. The randomization test permitted observation of the difference between the gain scores to determine if the obtained difference was within the region of rejection.

The null hypothesis suggested that the scores from the study were derived from the same population. That is,

---

[2]Sidney Siegel, Nonparametric Statistics for the Behavioral Sciences (New York: McGraw-Hill Book Company, Inc., 1956), pp. 152-156; see also Lincoln E. Moses, "Non-Parametric Statistics for Psychological Research," Psychological Bulletin, XLIX (1952), 122-143.

[3]J.P. Dickinson, Statistics for Business, Finance, and Accounting (Estover, Plymouth: Macdonald and Evans Limited, 1976), ch. 11-12.

[4]Frederick C. Mills, Introduction to Statistics (New York: Henry Holt and Company, 1956), ch. 9-10.

[5]John Neter and others, Fundamental Statistics for Business and Economics (4th ed.; Boston: Allyn and Bacon, Inc., 1973), ch. 22-24, 29.

chance would explain any numerical difference between group means. With five scores in each of two groups there was a possibility of 252 equally likely combinations of the scores between the two groups. To derive this number, one is required to use factorials as 10.9.8.7.6.5.4.3.2.1 which is divided by 5.4.3.2.1.5.4.3.2.1 .

Only once would the five smallest scores be found in one group while the five largest scores were found in the other group. Such a condition for a two-tailed test of the hypothesis would be, 2/252=1/126, of the possible combinations. If found, we would reject the null hypothesis at the .008 level of significance.

A level of significance of .05 permitted determination of the number of cases of extreme differences that composed the region of rejection for this study, (.05 times 252=12.6). As no direction of difference was predicted, a two-tailed test of the null hypothesis was proper.

Therefore, six of the most extreme possible outcomes in both directions were identified. If any of the combinations of extreme scores that constituted the region of rejection matched the results of the study, then the null hypothesis was rejected.

This study was designed to permit examination of the observations as a time series. In effect, to determine the

functional relationship of time and the observations, the study specified a task to be completed (a change in testing behavior), then examined the time required to complete the task through simple regression analysis.

As the first step of the analysis, a graph of the observed indices was presented, but, as convention requires, without lines connecting the observations. In a time series, time is the independent variable, and the observations of the indices are the dependent variables.

The shape of the above scatter diagram suggested the appropriate regression model for the time series. Further, the pictorial permitted determination of the direction of relationship, as inverse or direct.

The functional relationship may be expressed, in its simplest form, as Y = X, when one unit change in X predicts a one unit change in Y. Therefore, when the rate of change for Y is constant, the relationship may be expressed Y = a + bX, where

Y = Estimated mean value of the dependent variable
X = Time, or the independent variable
a = Point of intersection of the best fitting straight line with the Y axis, or the intercept
b = The rate of change of Y per unit change in X, the slope.

When actual, or observed, values of Y do not equal the estimated values of Y, the deviation from the estimated mean for each value of Y is a residual. When a line is

plotted for estimated values of Y that gives the least sum when the residuals are squared and summed, the best fitting straight line has been obtained. The statistical technique for this is known as the Method of Least Squares.

The purpose of regression analysis for this study was to determine the functional relationship between time and the dependent variables of Reliability Index, Item Difficulty Index, and Item Discrimination Index. Using such a technique to show a trend in a time series was reported favorably in the literature.[6]

There is an accepted sequence of operations that compose the technique of simple regression analysis when reporting association between variables. Merrill,[7] Mills,[8] and Dickinson[9] agreed that the first step is to present a scatter diagram; then to determine a least squares equation; secure a measure of the error involved in predicting Y,

---

[6]Dickinson, op. cit., p. 268; see also Mills, op. cit., p. 298.

[7]William C. Merrill and Karla Fox, Introduction to Economic Statistics (New York: John Wiley and Sons, Inc., 1970), p. 359.

[8]Mills, op. cit., pp. 214-240.

[9]Dickinson, op. cit., pp. 236-255.

(the Standard Error of the Estimate); compute the Coefficient of Correlation which measures the degree of relationship between the variables, (how the Standard Error of the Estimate compares to the Standard Deviation); and finally, to compute the Coefficient of Determination which, as the square of the Coefficient of Correlation, reports the amount of the total variance in Y explained by the model.

Chou stated that, although time series data may fail to meet necessary assumptions, it is no worse a method than the use of any other trend line. Further, such a technique is objective, and it is the one that statisticians have become accustomed to using.[10] The use of the least squares method is accepted for forecasting if we assume that the factors influencing the original data will be present at a future time, according to Chou[11] and Mandel.[12] Blair[13] found extrapolation acceptable if the forecast were not extended too far into the future.

---

[10]Ya-Lun Chou, Applied Business and Economic Statistics (New York: Holt, Rinehart and Winston, 1963), p. 504.

[11]Ibid., pp. 508-509.

[12]B.J. Mandel, Statistics for Management: A Simplified Introduction to Statistics (Baltimore: Dangary Publishing Co., 1956), p. 287.

[13]Morris M. Blair, Elementary Statistics With General Applications (New York: Henry Holt and Company, 1944), p. 392.

Wolberg found that "the distributions do not necessarily have to be normal" when we apply the method of least squares.[14] He suggested that his assumption is valid "as long as 'systematic errors' do not affect the measurements."[15] Four authors suggested that, although the method of least squares was not entirely appropriate for analysis of time series, its use was defensible in that empirical proof has been offered as to its practicality and expediency.[16]

The trend examined through this project was the average increase, or decrease, of the dependent variables within the identified time period of six events. Blair suggested that, "Time series trends are only a special case under general regression."[17] Further, he wrote that for computing trends from data over five to twelve years, or as in this project, time periods, the most satisfactory device is the straight-line least squares method.[18]

---

[14]John R. Wolberg, Prediction Analysis (Princeton: D. Van Nostrand Company, Inc., 1967), p. 30.

[15]Ibid., p. 31.

[16]Neter, op. cit., pp. 651; see also Mills, op. cit., pp. 298-299.

[17]Blair, op. cit., p. 370.

[18]Ibid., p. 378.

Dickinson noted that "over a relatively short period of time the assumption of linearity is in general a reasonable one."[19] He demonstrated, in detail as his Figure 71, "some possible non-linear relationships between x and y,"[20] and showed that segments of the lines were linear over short ranges. Chou also concluded that "we can often use linear regression to advantage for many problem situations within narrow ranges of X."[21]

This ability for a narrow range of X to be considered linear is also a limitation when regression values are used for prediction. Most texts contained warning about attempting to predict beyond the fixed range of X as the unknown values of Y could contribute to the relationship being non-linear. Therefore, use of the regression line to predict was considered to be a matter of good judgment.

---

[19] Dickinson, op. cit., p. 269.

[20] Ibid., p. 237.

[21] Chou, op. cit., p. 424.

IX. THE TEST ANALYSIS-FEEDBACK SYSTEM

This section contains elements that may add to the understanding of the treatment, and clarify some questions remaining about the data collection techniques. If the reader will first examine appendices A and C, he or she should be able to design and implement the system upon completion of the reading of this section.

The output appears as Appendix D. This output can be obtained by having answer decks compiled as instructed in Appendix A, or by keypunching the answers on one unit record card, as was the case when the control group data was prepared for analysis by the system. For the control data, there was one unit record card per student.

An observer who could be present for all processes would see the following behaviors:

(1) The teacher would design items for the test of the type that could be analyzed by the system.

(2) The test would be administered to the students.

(3) During or after the test, the students would be given the envelopes containing their computer cards.

(4) During or after the test, the students would follow the directions given on one of the envelopes for proper compiling of an answer deck.

(5) The envelopes would be turned in to the teacher who would take them, along with a teacher answer deck, to an agreed upon pickup area.

(6) A phone call to the number given in the booklet would result in the cards being picked up by the end of the school day. They would be taken to the computing center at The University of Tennessee, removed from the envelopes, stacked behind a Fortran program, and read into the computer.

(7) The feedback would be returned by the next school day. The envelopes containing the cards would be stored until the next test.

The Test Analysis-Feedback might more properly be considered a package rather than a single concept for experimentation. This is not seen as good or bad in an experimental sense, but does obviously require enough of a description of the system so it can be replicated by another researcher, and perhaps for future new explorations to be able to take apart the system for study.

CHAPTER IV

PRESENTATION OF THE DATA

I. INTRODUCTION

This chapter contains the introduction, presentation and analysis of the data used to examine the effect of the implementation of the Test Analysis Feedback System on the testing behavior of teachers, and a summary. The data were obtained from a population of ten teachers. Five of the teachers were randomly assigned to the experimental group, and five were randomly assigned to the control group.

The data were composed of a reliability index, an item difficulty index, and an item discrimination index. The data were obtained from the output of the Test Analysis-Feedback System.

The randomization test for two independent samples was chosen to test the significance of the difference between the gains of the two independent samples. The null hypothesis was rejected at the .05 level of significance. Simple regression analysis of the data was calculated to determine the functional relationship of an independent variable, time, and the dependent variables listed above.

II. PRESENTATION AND ANALYSIS

To give direction to the examination of the effect of a Test Analysis-Feedback System on the testing behavior of teachers, three hypotheses were formed. The first hypothesis was that the means of the pre-treatment to post-treatment gain scores would not significantly differ for the reliability indices of the experimental and the control groups.

For the reliability indices, as the observed combination of gain scores was not one of the scores within the region of rejection, the decision was to fail to reject the hypothesis at the .05 level of significance. The exact probability of the observed combination of scores was reported as, $p=.56$, for a two-tailed test.

The observed scores of the reliability indices for the members of the two groups, along with the number of students, and the number of test items, used as a base for determining each observation composed Table I. Figure 1, Appendix E, a graphical presentation, permitted visual inspection for any unique event that did not regular change.

The twelve combinations of scores that compose the region of rejection are presented as Table II. The observed gain scores were also identified within the table.

The second hypothesis proposed that the means of the pre-treatment to post-treatment gain scores would not

TABLE I

OBSERVATIONS OF THE RELIABILITY INDEX BY SUBJECT BY TEST

Group[a] (Subject)	Reliability Index for Test						Number of Students for Test						Number of Items per Test					
	01	02	03	04	05	06	01	02	03	04	05	06	01	02	03	04	05	06
E (1)	66	93	80	85	81	93	24	23	21	19	18	24	10	29	25	20	20	20
E (2)	78	91	81	85	68	80	71	70	71	64	63	66	25	35	30	30	25	17
E (3)	48	82	61	81	81	29	32	36	35	36	36	35	15	12	10	30	10	20
E (4)	50	38	67	42	34	67	21	19	18	17	19	16	14	20	20	20	20	20
E (5)	74	79	72	73	82	82	16	17	16	17	17	18	35	45	25	30	20	20
C (6)	91	72	86	80	51	93	20	19	16	19	18	18	36	25	50	25	18	40
C (7)	60	83	61	68	79	93	27	25	22	24	28	21	33	25	28	25	34	24
C (8)	92	84	82	95	92	74	22	21	22	24	24	24	26	17	20	50	43	34
C (9)	94	55	79	85	47	76	20	14	14	14	14	16	21	22	18	10	16	18
C (10)	89	24	72	92	84	88	26	22	22	21	23	24	16	10	15	50	27	41

[a]C = Member of the control group.
E = Member of the experimental group.

TABLE II

THE SIX MOST EXTREME POSSIBLE OUTCOMES IN EITHER DIRECTION FOR THE INDICES OF RELIABILITY

(The twelve possible combinations constitute the two-tailed region of rejection at the .05 level of significance.)

	Experimental Group					Sum	Control Group					Sum
Possible	-19	-18	-18	-01	002	-54	002	008	017	027	033	087
	002	008	017	027	033	087	-19	-18	-18	-01	002	-54
Pre-Treatment	-19	-18	-18	-01	002	-54	002	008	017	027	033	087
	002	008	017	027	033	087	-19	-18	-18	-01	002	-54
To	-19	-18	-18	002	002	-51	-01	008	017	027	033	084
	-01	008	017	027	033	084	-19	-18	-18	002	002	-51
Post-Treatment	-19	-18	-18	-01	008	-48	002	002	017	027	033	081
	002	002	017	027	033	081	-19	-18	-18	-01	008	-48
Gain Scores	-19	-18	-18	002	008	-45	-01	002	017	027	033	078
	-01	002	017	027	033	078	-19	-18	-18	002	008	-45
	-19	-18	-18	002	008	-45	-01	002	017	027	033	078
	-01	002	017	027	033	078	-19	-18	-18	002	008	-45
Observed Gain Scores	-19	002	008	017	027*	035	-18	-18	-01	002	033	-002

*.56 level of significance.

significantly differ for the difficulty indices of the experimental and the control groups. As the observed combination of gain scores of the item difficulty indices was not found within the region of rejection, the decision was to fail to reject the null hypothesis at the .05 level of significance. The exact probability was reported as, $p=.80$, for a two-tailed test.

The observed scores of the item difficulty indices for the experimental group, and for the control group compose Table III. Six observations, along with the number of students, and the number of test items, used as a base for determining each observation, were reported. Figure 2, Appendix E, permitted visual inspection for any unique event. Table IV reports the twelve combinations of scores that compose the region of rejection at the .05 level of significance.

It was also hypothesized that the means of the pre-treatment to post-treatment gain scores would not significantly differ for the discrimination indices of the experimental and control groups. It was determined that the same decision that was made for the item difficulty, and the reliability, variables should also be made for the item discrimination variable. That was, at the .05 level of significance, the appropriate decision was to fail to reject the hypothesis of no difference between the means of

TABLE III

OBSERVATIONS OF THE DIFFICULTY INDEX BY SUBJECT BY TEST

Group[a] (Subject)	Difficulty Index for Test						Number of Students for Test						Number of Items per Test					
	01	02	03	04	05	06	01	02	03	04	05	06	01	02	03	04	05	06
E (1)	25	37	17	26	31	43	24	23	21	19	18	24	10	29	25	20	20	20
E (2)	22	31	27	21	17	28	71	70	71	64	63	66	25	35	30	30	25	17
E (3)	34	37	34	14	31	29	32	36	35	36	36	35	15	12	10	30	10	20
E (4)	41	40	51	37	27	31	21	19	18	17	19	16	14	20	20	20	20	20
E (5)	27	29	22	22	29	24	16	17	16	17	17	18	35	45	25	30	20	20
C (6)	19	33	19	37	47	44	20	19	16	19	18	18	36	25	50	25	18	40
C (7)	23	27	15	24	07	25	27	25	22	24	28	21	33	25	28	25	34	24
C (8)	18	12	17	26	08	08	22	21	22	24	24	24	26	17	20	50	43	34
C (9)	33	09	32	25	10	11	20	14	14	14	14	16	21	22	18	10	16	18
C (10)	33	45	26	31	25	29	26	22	22	21	23	24	16	10	15	50	27	41

[a]C = Member of the control group.
E = Member of the experimental group.

TABLE IV

THE SIX MOST EXTREME POSSIBLE OUTCOMES IN EITHER DIRECTION FOR THE INDICES OF DIFFICULTY

(The twelve possible combinations constitute the two-tailed region of rejection at the .05 level of significance.)

	Experimental Group					Sum	Control Group					Sum
Possible	-22	-10	-10	-05	-04	-51	-03	002	006	018	025	048
	-03	002	006	018	025	048	-22	-10	-10	-05	-04	-51
Pre-Treatment	-22	-10	-10	-05	-03	-50	-04	002	006	018	025	047
	-04	002	006	018	025	047	-22	-10	-10	-05	-03	-50
To	-22	-10	-10	-04	-03	-49	-05	002	006	018	025	046
	-05	002	006	018	025	046	-22	-10	-10	-04	-03	-49
Post-Treatment	-22	-10	-10	-05	002	-45	-04	-03	006	018	025	042
	-04	-03	006	018	025	042	-22	-10	-10	-05	002	-45
Gain Scores	-22	-10	-10	-04	002	-44	-05	-03	006	018	025	041
	-05	-03	006	018	025	041	-22	-10	-10	-04	002	-44
	-22	-10	-05	-04	-03	-44	-10	002	006	018	025	041
	-10	002	006	018	025	041	-22	-10	-05	-04	-03	-44
Observed Gain Scores	-10	-05	-03	006	018*	006	-22	-10	-04	002	025	-09

*.80 level of significance.

the gain scores. The exact probability was reported as, $p=.30$, for a two-tailed test.

The observed scores of the item discrimination variable for the experimental, and for the control groups, along with the number of students and test items used as a base for determining each observation compose Table V. Visual inspection of the data were possible in Figure 3, Appendix E. Table VI reported the 12 combinations of scores that were the region of rejection at the .05 level of significance.

The optimum value of the reliability index was reported as greater than 69 for teacher made tests. The optimum value for the difficulty index was given as a range from 25 to, and including, 75. The optimum value for the discrimination index was considered to be from 40 to, and including, 100. Figures 4, 5, and 6, Appendix E, were plots of the observed scores for each participant for each index, but with the optimum values for the index shown.

The average, as a measure of central tendency, has a disadvantage in that it is responsive to extreme scores. One deviant observation can raise, or lower, an overall mean considerably. Therefore, Tables VII and VIII, with Figures 7 and 8, Appendix E, were included to report the percent of those items that attained the optimum value for item difficulty, and item discrimination, respectively.

TABLE V

OBSERVATIONS OF THE DISCRIMINATION INDEX BY SUBJECT BY TEST

Group[a] (Subject)	Discrimination Index for Test						Number of Students for Test						Number of Items per Test					
	01	02	03	04	05	06	01	02	03	04	05	06	01	02	03	04	05	06
E (1)	50	56	29	48	44	69	24	23	21	19	18	24	10	29	25	20	20	20
E (2)	31	50	34	33	24	49	71	70	71	64	63	66	25	35	30	30	25	17
E (3)	38	58	55	25	61	27	32	36	35	36	36	35	15	12	10	30	10	20
E (4)	37	29	40	29	25	39	21	19	18	17	19	16	14	20	20	20	20	20
E (5)	29	29	33	26	45	48	16	17	16	17	17	18	35	45	25	30	20	20
C (6)	32	34	32	46	32	58	20	19	16	19	18	18	36	25	50	25	18	40
C (7)	23	41	22	33	15	45	27	25	22	24	28	21	33	25	28	25	34	24
C (8)	35	25	34	51	16	16	22	21	22	24	24	24	26	17	20	50	43	34
C (9)	66	18	44	50	20	22	20	14	14	14	14	16	21	22	18	10	16	18
C (10)	62	33	40	45	41	41	26	22	22	21	23	24	16	10	15	50	27	41

[a]C = Member of the control group.
E = Member of the experimental group.

TABLE VI

THE SIX MOST EXTREME POSSIBLE OUTCOMES IN EITHER DIRECTION FOR THE INDICES OF DISCRIMINATION

(The twelve possible combinations constitute the two-tailed region of rejection at the .05 level of significance.)

	Experimental Group					Sum	Control Group					Sum
Possible	-44	-21	-19	-11	002	-93	018	019	019	022	026	104
	018	019	019	022	026	104	-44	-21	-19	-11	002	-93
Pre-Treatment	-44	-21	-19	-11	018	-77	002	019	019	022	026	088
	002	019	019	022	026	088	-44	-21	-19	-11	018	-77
To	-44	-21	-19	-11	019	-76	002	018	019	022	026	087
	002	018	019	022	026	087	-44	-21	-19	-11	019	-76
Post-Treatment	-44	-21	-19	-11	019	-76	002	018	019	022	026	087
	002	018	019	022	026	087	-44	-21	-19	-11	019	-76
Gain Scores	-44	-21	-19	-11	022	-73	002	018	019	019	026	084
	002	018	019	019	026	084	-44	-21	-19	-11	022	-73
	-44	-21	-19	-11	026	-69	002	018	019	019	022	080
	002	018	019	019	022	080	-44	-21	-19	-11	026	-69
Observed Gain Scores	-11	002	018	019	019*	047	-44	-21	-19	022	026	-36

*.30 level of significance.

TABLE VII

PERCENT OF ITEMS ACHIEVING OPTIMUM VALUES FOR ITEM DIFFICULTY

Group[a] (Subject)	Percent of Items Achieving Optimum Value for Test						Number of Students for Test						Number of Items per Test					
	01	02	03	04	05	06	01	02	03	04	05	06	01	02	03	04	05	06
E (1)	70	93	28	55	65	100	24	23	21	19	18	24	10	29	25	20	20	20
E (2)	40	71	40	40	32	47	71	70	71	64	63	66	25	35	30	30	25	17
E (3)	73	83	70	23	70	60	32	36	35	36	36	35	15	12	10	30	10	20
E (4)	93	65	90	70	45	70	21	19	18	17	19	16	14	20	20	20	20	20
E (5)	66	51	52	33	65	45	16	17	16	17	17	18	35	45	25	30	20	20
C (6)	22	76	40	72	61	85	20	19	16	19	18	18	36	25	50	25	18	40
C (7)	30	52	36	60	06	63	27	25	22	24	28	21	33	25	28	25	34	24
C (8)	42	29	30	64	30	06	22	21	22	24	24	24	26	17	20	50	43	34
C (9)	62	14	83	70	25	33	20	14	14	14	14	16	21	22	18	10	16	18
C (10)	69	90	47	80	56	68	26	22	22	21	23	24	16	10	15	50	27	41

[a]C = Member of the control group.
E = Member of the experimental group.

TABLE VIII

PERCENT OF ITEMS ACHIEVING OPTIMUM VALUES FOR ITEM DISCRIMINATION

Group[a] (Subject)	Percent of Items Achieving Optimum Value for Test						Number of Students for Test						Number of Items per Test					
	01	02	03	04	05	06	01	02	03	04	05	06	01	02	03	04	05	06
E (1)	70	76	20	60	65	90	24	23	21	19	18	24	10	29	25	20	20	20
E (2)	36	69	43	40	24	59	71	70	71	64	63	66	25	35	30	30	25	17
E (3)	60	83	60	27	100	45	32	36	35	36	36	35	15	12	10	30	10	20
E (4)	57	65	60	40	40	55	21	19	18	17	19	16	14	20	20	20	20	20
E (5)	40	40	44	30	65	75	16	17	16	17	17	18	35	45	25	30	20	20
C (6)	53	56	34	76	56	88	20	19	16	19	18	18	36	25	50	25	18	40
C (7)	24	60	32	36	06	50	27	25	22	24	28	21	33	25	28	25	34	24
C (8)	42	29	30	62	30	06	22	21	22	24	24	24	26	17	20	50	43	34
C (9)	81	14	72	70	25	33	20	14	14	14	14	16	21	22	18	10	16	18
C (10)	81	30	33	58	37	49	26	22	22	21	23	24	16	10	15	50	27	41

[a]C = Member of the control group.
E = Member of the experimental group.

Plots representing the difficulty indices, in Figure 7 and 2, Appendix E, apparently demonstrated that the index and optimum value scores tended to vary together in degree, and direction. The percentage of items meeting the criterion of optimum difficulty value apparently tended to reflect the observed values of the item difficulty index for individual members of the experimental, and the control, groups.

The plots representing the discrimination indices, Figures 8 and 3, Appendix E, apparently demonstrated that the scores tended to vary together for both groups in both degree and direction. That is, the percentage of items within the optimum discrimination range apparently tended to reflect the observed values of the item discrimination index for the individual members of both the experimental and the control groups. The quantity of items attaining optimum value did not increase, nor decrease, independently of their indices.

To provide further insight into the relationships between the variables, the following descriptions, and tentative analysis, of the data were provided. As the first step in the examination of the population parameters as data in a time series, Figure 9, Appendix E, a scatter diagram of the reliability indices for the experimental group was constructed.

Time, the independent variable was plotted on the horizontal axis, in Figure 9, Appendix E, while the values

that could be taken by the dependent variable, the reliability indices, were plotted on the vertical axis. The shape of the scatter diagram in Figure 9, Appendix E, suggested that the appropriate regression model could be linear. The plots of Figure 9, Appendix E, appeared to be too random to determine whether the relationship of the two variables was inverse, or direct.

The scatter diagram of the reliability indices for the control group is Figure 10, Appendix E. As in the previous figure, the pictorial did not suggest the direction of the relationship for the variables. The apparent regression model was linear.

The scatter diagram of the difficulty indices for the experimental group is Figure 11, Appendix E, while Figure 12, Appendix E, is the scatter diagram of the difficulty indices for the control group. In both cases, the apparent regression model was linear. In neither case was it apparent as to whether the relationship of the two variables was inverse or direct.

The scatter diagrams for the item discrimination indices for the experimental, and the control groups are Figures 13 and 14, Appendix E, respectively. Their apparent regression models were linear. The relationship, as inverse or direct, of their variables could not be determined from visual inspection.

The slope, representing the mangnitude of the change in the dependent variable for a given change in the independent variable, and the intercept, two values that determine the prediction for the estimated values of the dependent variables are included as Table IX. Further, the table contains the measure of error in predicting the dependent variable, the standard error of the estimate. The coefficient of correlation, (r), the measure of relationship between the variables, or how the standard error of the estimate compares to the standard deviation of the dependent variable was included in the table.

The significance of r, the mean of the dependent variable, its standard deviation, and the number of observations were reported as Table IX. Also included was the coefficient of determination, (r^2), which was used to suggest the amount of the total variance in the dependent variable explained by the regression model. The scatter diagrams, and the statistics associated with the simple regression, were the output of the system of computer programs known as the Statistical Package for the Social Sciences (SPSS).

The following analysis may be followed in Table IX, and Figures 9 through 14, Appendix E. First, the coefficient of correlation supported the tentative conclusion of little relationship between time and the dependent variables. Regression theory proposes that the better the

TABLE IX

POPULATION PARAMETERS FOR TREND ANALYSIS

Index (Group)	Slope	Intercept	See	r	Sig	r^2	$\overline{Y}$	Sd	N
Reliability (Experimental)	0.39429	69.38667	17.9	.04	.42	.002	71	18	30
Reliability (Control)	0.77143	74.66667	17.0	.08	.34	.006	77	17	30
Difficulty (Experimental	-0.67429	31.82667	8.3	-.14	.23	.02	29	8	30
Difficulty (Control)	-0.560	25.8933	11.4	-.09	.33	.007	24	11	30
Discrimination (Experimental)	0.77714	36.94667	12.5	.11	.28	.01	40	14	30
Discrimination (Control	-1.18857	39.89333	13.8	-.15	.21	.02	36	14	30

See = Standard error of the estimate.
r = Pearson product-moment correlation coefficient.
Sig = Significance of r
r^2 = Coefficient of determination
$\overline{Y}$ = Mean of the index values
Sd = Standard deviation for Y
N = Number of indices

fit of the data on the trend line, the larger will be the reported magnitude of r.

Further, r^2, the coefficient of determination, for each of the six indices was extremely small. From this observation it was tentatively proposed that very little of the variance of the scores of the dependent variables was explained by the independent variable, time.

Examination of the scatter diagrams confirmed the extremely low r^2. The plots for the reliability indices, the item difficulty indices, and the item discrimination indices appeared to be spread around the trend line in an almost random fashion for both the experimental group, and for the control group. From observation of such plots, and from the information in Table IX, it appeared that any prediction of the value of the dependent variables based on time would be, for all practical purposes, worthless.

The trend line for the reliability indices of the experimental group, and the trend line for the reliability indices of the control group, as shown in Figure 9, Appendix E, and Figure 10, Appendix E, appeared to vary together. This visual interpretation was supported by the values of the slope, and the values of the intercept, reported in Table IX. Table IX and Figure 11, Appendix E, and Figure 12, Appendix E, demonstrated the same condition, apparently, for the difficulty variable.

Although similar intercept values for the discrimination variable, (37 to 40) were shown in Table IX, the slope value for the control group represented an inverse relationship between time and the dependent variable, while the slope of the best fitting line for the experimental group signified a direct relationship. That is, over time, the item discrimination index for the experimental group was inclining, while the index for the control group was declining. In both cases, the coefficient of determination, (r^2), warned that very little of the variance of the dependent variable was explained by the independent variable, time.

One of the problems of trend analysis is the determination as to which part of the total period covered by the data should be included in computing the trend line. For example, if any observation were a unique event, then inclusion of that observation in the analysis could cause an increase, or decrease, in the resulting trend line.

When attempting to obtain a trend line with greater magnitude, convention permits an analyst to select a starting point within the data. The starting point that is selected usually excludes any possibly unique events.

For this study, a series of time used in calculating such a new trend line was adjusted to include observations of time periods three through six. Table X was formed from the values derived by analysis of observations three through

TABLE X

POPULATION PARAMETERS FOR TREND ANALYSIS (MODIFIED)

Index (Group)	Slope	Intercept	See	r	Sig	r^2	$\overline{Y}$	Sd	N
Reliability (Experimental)	-1.0	75.70	17.9	-.06	.78	.004	71	17	20
Reliability (Control)	1.30	73.0	13.9	.11	.65	.01	79	14	20
Difficulty (Experimental)	0.540	25.620	9.0	.07	.77	.005	28	9	20
Difficulty (Control)	-0.440	25.280	11.9	-.04	.86	.002	23	12	20
Discrimination (Experimental)	3.220	24.660	12.8	.29	.22	.08	39	13	20
Discrimination (Control)	-1.420	41.540	13.3	-.13	.60	.02	35	13	20

See = Standard error of the estimate.
r = Pearson product-moment correlation coefficient.
Sig = Significance of r.
r^2 = Coefficient of Determination
$\overline{Y}$ = Mean of the index values
Sd = Standard deviation for Y
N = Number of indices

six. Disregarding plots for time one and time two, Figures 9 through 14, Appendix E, were viewed as the scatter diagrams that would be derived by plotting observations three through six.

It was tentatively determined from comparison of the parameters presented in Table IX, page 79 and Table X, page 82, that the slope and intercept values for the control group did not change in magnitude, nor direction. For the experimental group, the direction of the slope apparently reversed for reliability and difficulty indices. It appeared that, while the direction of the slope for the discrimination variable remained the same, the magnitude of the slope did increase in the desired direction.

Therefore, it is possible that observation two for the discrimination variable for the experimental group may have been a unique event that may be explained by the treatment effect. Recognizing that any two given sets of data would result in the calculation of different intercept, and slope values, and facing the realities of the r^2, a definite causal relationship was not suggested.

III. SUMMARY

The decision was to fail to reject the hypotheses of no difference between the means of the gain scores for

the variables of reliability, difficulty, and discrimination. The exact probabilities were reported as .56, .80, and .30, for reliability, difficulty, and discrimination, respectively. The analysis of the time series data produced estimated values for the dependent variables that had relatively large variances.

CHAPTER V

SUMMARY, CONCLUSIONS, DISCUSSION, AND RECOMMENDATIONS

I. SUMMARY

The problem of this study was the need for more effective tools for a supervisory behavior system to utilize to influence the classroom testing behavior of teachers. The purpose of the study was to examine the effect of the Test Analysis-Feedback System on the testing behavior of teachers, as a first step in the development of a supervisory support system dealing with student evaluation.

To give direction to the study it was hypothesized that there would be no difference between the pre-treatment to post-treatment gain scores for the dependent variables of reliability, item difficulty, and item discrimination. The null hypothesis was rejected at the .05 level.

The population was composed of ten teachers. To form an experimental group and a control group, members of the population were assigned randomly to one of the two treatment conditions. Only the experimental group was treated with information from the Test Analysis-Feedback System. That information was described as Appendix A, and as Appendix C.

Data were collected from copies of the output of the system. The reliability index, the difficulty index,

and the discrimination index were determined.

The randomization test for two independent samples was chosen to determine the significance of the difference between the means of the gain scores of the two independent samples. Observations for analysis included the reliability index, the difficulty index, and the discrimination index. Test one was the pre-treatment observation. Test six was the post-treatment observation.

To determine the functional relationship between time and the dependent variables of test reliability, item difficulty, and item discrimination, a simple regression analysis was performed.

The following findings were made. At the .05 level of significance, there was no difference between the means of the pre-treatment to post-treatment gain scores of the two groups. Therefore, the decisions were to fail to reject the hypotheses of no difference for the variable of reliability, difficulty, and discrimination. The exact probaabilities were reported as .80, .56, and .30, for difficulty, reliability, and discrimination, respectively.

Tentatively, after examination of the population parameters that described the time series data, it was found that, although there were differences in the slopes of the trend lines for any given dependent variable, the strength of association was weak. In an effort to obtain lines with

greater magnitude, the starting point of the time series was modified. By comparison of the parameters of the original and the modified time series, it was shown that the slope and intercept values for the control group did not change in magnitude, nor in direction, with the creation of new data.

For the experimental group, the direction of the slope reversed for the reliability indices, and for the difficulty indices. Although the direction of the slope remained the same for the discrimination indices, the magnitude of the slope apparently did increase in the desired direction.

II. THE CONCLUSIONS

The concept that was isolated for examination was the implementation of the Test Analysis-Feedback System within a testing program. As the experimental group was given information that was not made available to the control group, it was logical to consider that there was a difference between the knowledge of members of the two groups concerning their tests.

It was concluded that the Text Analysis-Feedback System did not have measureable effects on the testing behavior of members of the experimental group. That is, the knowledge was not translated into effective action.

The Test Analysis-Feedback System did not act to narrow the gap between an implicit goal of educational organizations, more effectiveness in testing, and the goals of the individual. In other words, increased information about a procedure being used to evaluate students, did not result in a significant change in behavior.

Analysis of the data did not, therefore, support a conclusion that feedback directly to the individual concerned could, by itself, influence the behavior of an individual. This was not seen to be the same as concluding that the analysis supported that a supervisory behavior system is necessary to effect change.

Further, the quality indices of the observed dependent variables were not cosistent over time, with or without, the information from the Test Analysis-Feedback System. That is, for both groups, the index values tended to fluctuate. Apparently, neither of the groups was provided with the necessary support system to improve the quality of their testing behavior, or to insure its consistency.

Finally, individuals were able to operationalize the concept of the Test Analysis-Feedback System. The structure of the system did fit into the classroom testing procedures.

III. DISCUSSION

Within the period of time covered by the experiment, it was expected that a change in the testing behavior of the teachers would be measureable, either as a cumulative effect in the sixth observation, or gradually over time. Neither of these was supported by the findings.

Design considerations limited the use of quantitative analysis for adjacent time periods. Therefore, to provide insight, and to present as much information as possible, this discussion was presented.

Considering the literature, and after examining the individual scatter diagrams and data within the tables, it was conjectured that the second observations of the dependent variables for the experimental group were reflecting the treatment effect. Of the 25 values that composed the data for each group, (reliability indices, item difficulty indices, item discrimination indices, the percent of items that had achieved optimum item difficulty, and the percent of items that had achieved optimum item discrimination) 18 of the scores of the experimental group showed an increase over the values of their first observations. Eleven of the scores of the control group showed an increase. The significance between the observed proportional distribution for the experimental group and what might have resulted from chance was computed using the chi-square test. A

value of 4.84 constituted a significant departure from what would have been expected by chance. That is, it could have been concluded that the observed difference in the direction of change was significant at the .05 level. Therefore, the null hypothesis could have been rejected.

From the qualitative look at the data of the individual participants, it was conjectured that the Test Analysis-Feedback System did contribute to unfreezing between observations one and two, and did contribute to moving the testing behavior of the teachers to a new level. Unfortunately, it appeared that the Test Analysis-Feedback System was unable to contribute to refreezing the behavior at the new level. That is, the intervention of the feedback seemed to have been a catalyst for change, as intended, but it also seemed to lack the utility to maintain the new level.

Apparently, the new behavior did not fit in with the established procedures, norms, and avalues of the subsystem of testing. As a result, not enough support was received by the teacher for the factors that contributed to the equilibrium of the testing behavior system to be upset over the time period of the experiment.

It may have been that the goal of improved testing behavior was seen as self-enhancing by the individuals, that they did develop a high level of interest, and that

they did interact with the system, but the presumed change did not seem to be sustained. This apparent failure of the system to continue to improve the quality of the testing behavior of the individuals may have reflected the need for some other intervention after the first feedback.

In retrospect, the tenability of assuming that the quantity of information presented by the Test Analysis-Feedback System, and the format in which the information was presented, were optimums, may be questionable. The charateristics of the knowledge may have been better presented with different processes, strategies, and amount of information.

For example, the format may not have stuck closely to specifics that had meaning to the experiences of the teachers in the experimental group. Further, even if there were no question of the format of presentation, the quantity of information may have ranged from too little to too much.

Another basic assumption was that the information that was presented by the system would be understandable by the teachers, where, in fact, an alternative condition may have been operating. A more tenable assumption may have been that being a teacher does not automatically predetermine the level of an individuals ability to effectively use the Test Analysis-Feedback System. Simply put, the competency level, or sense of competence for any or all participants, may have been minimal.

More formal and relevant training, or reeducation, may have needed to have been assured, rather than to have been assumed. There are a number of techniques which permit peers, supervisors, or outside consultants to intervene in the feedback process.

It is possible that the lack of measureable change reflected the influence of a number of other assumptions that are not tenable. Prominent among these was the assumption that the perceived relative advantage of the Test Analysis-Feedback System over the existing practices would insure that sufficient resources would be utilized to improve testing behavior. On the contrary, the necessary felt dissatisfaction with the testing behavior may have not been present to support the validity of such an assumption about any, or all, of the members of the experimental groups.

Further, even if the system were accepted as advantageous, the cost of past experiences with educational innovations may have influenced the behavior of the teachers. Therefore, no desired change in behavior would have been measureable as no treatment from the system would have been received.

An alternative plausible condition may be offered for the assumption that the tests would be constructed by the teachers in the experimental group after the intervention of feedback. The teachers may have simply administered

questions from the accompanying teacher editions of the materials.

Finally, as it was assumed that the magnitude of the difference between the gain scores had utility as an indicator of change, a test was used which represented whether the gain scores of two groups differed in a measure of central tendency, the means. Conventional use of such a test permits two choices as to the possibility of chance accounting for the difference. The choices are to reject, or to fail to reject the hypothesis of no difference.

In this study, the differences were regarded as not significant from that which could be expected from chance as they had highly insignificant values (exact probabilities of .80, .56, and .30). One possible alternative would have been to permit the reader to select his own dichotomy by merely reporting the findings with their levels of significance.

In this study, a reader could then accept that, relatively, the analysis of the indices revealed that the discrimination variable reflected at least marginally significant effect on the treatment. With such a procedure, it would be possible to preserve information that might otherwise be lost.

IV. RECOMMENDATIONS

The following were the recommendations from the study. First, it was recommended that this study be replicated to further contribute to knowledge about the effect of the Test Analysis-Feedback System on the testing behavior of teachers. Careful consideration should be given to the possible influence of the assumptions of this study.

A second recommendation was that a model be developed that could direct research efforts into the use of data-based feedback methods. Such a model would suggest data collection strategies, techniques for data analysis, and the method of transmission of the data. That is, the model would give direction for examination of why a system of feedback worked or did not work, from the viewpoint of who should do what to whom, when, and how.

As the nature of the Test Analysis-Feedback System itself may have limited its treatment effect, it was recommended that the content of the treatment variable be examined to determine the contribution of the separate parts of the feedback document to the treatment effect. Such an examination should seek to answer to question as to what inputs have utility for a supervisory behavior system to use to unfreeze, move, and refreeze the testing behavior of teachers.

As the quality of the testing behavior for members of both groups lacked consistency, it was recommended that colleges should investigate their responsibility for improving the testing behavior of teachers while they are still education majors. Further, school administrators and supervisors should examine the quality of the testing behavior of the teachers already in service and, if necessary, should implement staff development activities that will aid in removing deficiencies.

The system may have areas of application other than improvement of the testing behavior of teachers. For example, it would be possible to obtain information that could contribute to the quality control of the curriculum. Reports could be generated that presented profiles of student achievement of stated objectives. Therefore, examination of the utility of the system for curriculum control was recommended.

BIBLIOGRAPHY

BIBLIOGRAPHY

Alderfer, Clayton P., and John Holbrook. "A New Design for Survey Feedback," Education and Urban Society, V, 4 (1972/1973), 437-464.

Allen, Paul M., and others. Teaching Self-Appraisal: A Way of Looking Over Your Own Shoulder. Worthington, Ohio: Charles A. Jones Publishing Company, 1970.

Anderson, James G. Bureaucracy in Education. Baltimore: The Johns Hopkins Press, 1968.

Annett, John. Feedback and Human Behavior. Baltimore: Penguin Books Inc., 1969.

Association for Supervision and Curriculum Development. Evaluation as Feedback and Guide, 1967 Yearbook. Washington, D.C.: The Association's Press, 1967.

Beer, Michael, and Edgar F. Huse. "A Systems Approach to Organizational Development," The Journal of Applied Behavioral Science, VIII, 1 (1972), 79-101.

Bennis, Warren G., Kenneth D. Benne, and Robert Chin (eds.). The Planning of Change. New York: Holt, Rinehart and Winston, 1962.

Blair, Morris M. Elementary Statistics with General Applications. New York: Henry Holt and Company, 1944.

Bloom, Benjamin S., and others. Handbook on Formative and Summative Evaluation of Student Learning. New York: McGraw-Hill Book Company, 1971.

Chou, Ya-Lun. Applied Business and Economic Statistics. New York: Holt, Rinehart and Winston, 1963.

Collins, Barry E., and Harold Guetzkow. A Social Psychology of Group Processes for Decision Making. New York: John Wiley & Sons, 1964.

Dagnon, Carol, and Dennis W. Spuck, "A Role for Computers in Individualizing Education--and It's Not Teaching," The Phi Delta Kappan, LVIII (February, 1977), 460-462.

Davis, Frederick B. "Item Analysis in Relation to Educational and Psychological Testing," Psychological Bulletin, XLIX, 2 (March, 1952), 97-118.

Dickinson, J.P. Statistics for Business, Finance, and Accounting. Estover, Plymouth: Macdonald and Evans Limited, 1976.

Ebel, Robert L. "Can Teachers Write Good True-False Test Items?," Journal of Educational Measurement, XII, 1 (Spring, 1975), 31-35.

_____. Essentials of Educational Measurement. Englewood Cliffs, N.J.: Prentice-Hall, Inc., 1972.

Educational Testing Service. Short-Cut Statistics for Teacher-Made Tests, Evaluation and Advisory Series. Princeton: Educational Testing Service, 1960.

Feyereisen, Kathryn V., and others. Supervision and Curriculum Renewal: A Systems Approach. New York: Appleton-Century-Crofts, 1970.

French, Wendell L., and Cecil H. Bell, Jr. Organizational Development: Behavioral Science Interventions for Organizational Improvement. Englewood Cliffs, N.J.: Prentice-Hall, Inc., 1973.

Gage, N.L. (ed.). Handbook of Research on Teaching. Chicago: Rand McNally & Co., 1963.

Gulliksen, Harold. "The Relation of Item Difficulty and Inter-Item Correlation to Test Variance and Reliability," Psychometrika, X, 2 (June, 1945), 79-91.

Harris, Chester W. (ed.). Encyclopedia of Educational Research. 3d ed. New York: Macmillan Company, 1960.

Heller, Frank A. "Group Feed-Back Analysis as a Change Agent," Human Relations, XXIII, 4, (August, 1970), 319-333.

Kelley, Truman L. "The Selection of Upper and Lower Groups for the Validation of Test Items," Journal of Educational Psychology, XXX, 1 (January, 1939), 17-24.

Kenney, James B., and R. Robert Rentz. Automation and Control of Public School Instructional Records. Itasca, Illinois: F.E. Peacock Publishers, Inc., 1970.

Klein, Stuart M., Allen I. Kraut, and Alan Wolfson. "Employee Reactions to Attitude Survey Feedback: A Study of the Impact of Structure and Process," Administrative Science Quarterly, XVI, 4 (December, 1971), 497-513.

Kolb, D.A., and A.L. Frohman. "An Organization Development Approach to Consulting," Sloan Management Review, XII (1970), 51-65.

Kuder, G.G., and M.C. Richardson. "The Theory of the Estimation of Test Reliability," Psychometrika, II (September, 1937), 151-160.

Lewin, Kurt. Field Theory in Social Science. New York: Harper, 1951.

Lewis, Gertrude M. The Evaluation of Teaching. Washington, D.C.: National Education Association Publications, 1966.

Lindquist, E.G. (ed.). Educational Measurement. Washington, D.C.: American Council on Education, 1951.

Lippey, Gerald (ed.). Computer-Assisted Test Construction. Englewood Cliffs, N.J.: Educational Technology Publications, 1974.

Mandel, B.J. Statistics for Management: A Simplified Introduction to Statistics. Baltimore: Dangary Publishing Co., 1956.

Margulies, Newton, and Anthony P. Raia. Conceptual Foundations of Organizational Development. New York: McGraw-Hill Book Company, 1978.

Maslow, A.H. "A Theory of Human Motivation," Psychological Review, L (July, 1943), 370-396.

Merrill, William C., and Karla Fox. Introduction to Economic Statistics. New York: John Wiley and Sons, Inc., 1970.

Mills, Frederick D. Introduction to Statistics. New York: Henry Holt and Company, 1956.

Moses, Lincoln E. "Non-Parametric Statistics for Psychological Research," Psychological Bulletin, XLIX (1952), 122-143.

Nadler, David A. Feedback and Organization Development: Using Data-Based Methods. Reading, Mass.: Addison-Wesley Publishing Company, 1977.

_____. "The Use of Feedback for Organizational Change: Promises and Pitfalls," Group and Organizational Studies, I, 2 (June, 1976), 177-184.

_____. Philip Mirvis, and Cortlandt Cammann. "The Ongoing Feedback System: Experimenting with a New Managerial Tool," Organizational Dynamics, IV, 4 (Spring, 1976), 63-80.

National Society for the Study of Education. The Dynamics of Instructional Groups: Sociopsychological Aspects of Teaching and Learning. Fifty-ninth Yearbook, Part II. Chicago: University of Chicago Press, 1960.

Neter, John, and others. Fundamental Statistics for Business and Economics. 4h ed. Boston: Allyn and Bacon, Inc., 1973.

Overly, Donald E., and others. The Middle School: Humanizing Education for Youth. Worthington, Ohio: Charles A. Jones Publishing Company, 1972.

Parsons, Talcott, and E.A. Shils. Toward a General Theory of Action. Cambridge, Mass.: Harvard University Press, 1971.

Remmers, H.H., and N.L. Gage. Educational Measurement and Evaluation, rev. ed. New York: Harper & Brothers, 1955.

Remmers, H.H., and others (eds.). Growth, Teaching, and Learning: A Book of Readings. New York: Harper and Brothers, 1957.

Schoer, Lowell A. Test Construction: A Programmed Guide. Boston: Allyn and Bacon, 1970.

Siegel, Sidney. Nonparametric Statistics for the Behavioral Sciences. New York: McGraw-Hill Book Company, Inc., 1956.

Van Til, William, and others. Modern Education for the Junior High School Years. New York: The Bobbs-Merrill Company, Inc., 1961.

Wolberg, John R. Prediction Analysis. Princeton: D. Van Nostrand Company, Inc., 1967.

Zajonc, Robert B. "The Effects of Feedback and Probability of Group Success on Individual and Group Performance," Human Relations, XV, 2 (May, 1962), 149-161.

APPENDICES

APPENDIX A

TEST ANALYSIS-FEEDBACK SYSTEM

Thank you for participating in the development of a flexible and inexpensive system that will provide you with information about your classroom testing.

You often have to spend hour upon hour of your valuable, and limited, time scoring tests, either at home, or during class when you could be teaching. Although this problem has long been recognized, we do not have much information as to what can be done about it.

But, one thing we do know is that any effort to discover how a system could help with classroom testing has to involve you, and it has to infringe very little on your already limited time.

We believe that together we can discover the techniques and information that need to be a part of any system that is designed to help you with the tedious task of student evaluation.

Canton Napier

The University of Tennessee

WHAT IS THE "TEST ANALYSIS-FEEDBACK SYSTEM"?

The System is a technique that requires very little effort (or time) for a teacher to obtain information about classroom testing. The System is a way for providing you with the following information about each test:

A raw score for each student.
(20 correct)

The percent of correct responses for each student.
(80 percent)

Each answer marked as correct (1), or incorrect (0).
(10111 00111 11111)

The class average.
(81.3 percent)

How easy the item was for all students.
(11.1 percent * * * * * *)

The number of responses to each alternative with the correct response marked with a *.
(First 16 Second 9* Third 2 Fourth 1)

The difficulty level of each item.
(59.9)

The discrimination level of each item.
(.43)

The reliability of the test.
(.78)

WHAT IS REQUIRED OF THE TEACHER?

We hope to keep your work to a minimum with the following procedures:

1. Supply us with a list of your students. (We need to type a "name card" for each.)

2. Explain the procedure to the students. (As explained in detail below, and on their envelopes, they make a stack of cards that show their answers to your questions.)

3. Make a stack of teacher answers for the test.

4. Place the decks of cards where they can be picked up.

5. Pick up your test analysis for use in your classroom.

Except at your request, data will not be identified by teacher or student. Further, no use will be made of the data other than to evaluate the Test Analysis-Feedback System.

WHAT INSTRUCTIONS APPEAR ON THE STUDENT MATERIALS?

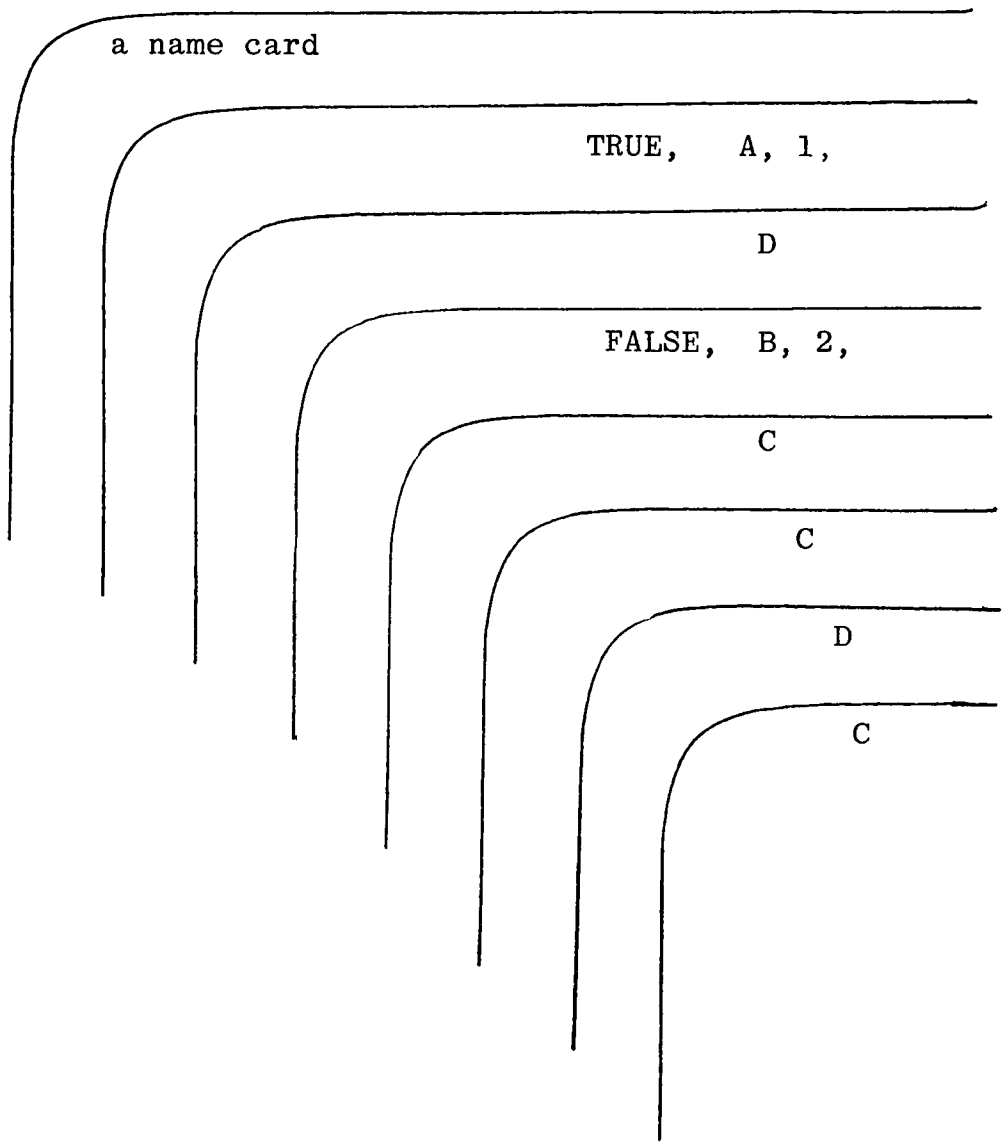
(At your convenience, during or after the test, give each student the two envelopes that store the computer cards.)

One of the envelopes contains instructions to:

1. Empty both envelopes.
2. Place your name card in this envelope so the name appears in the upper left.
3. Count out the number of answer cards you will need for the test. Put the remaining cards in your envelope marked STORAGE ONLY--TO BE USED FOR NEXT TEST.
4. To show your choice for an answer, start by placing a card on top of your name card to show your first answer. Turn the card so that your answer appears in the ANSWER SPACE. (Remember, all you have to do is turn the card to show the answer that you want.)
5. Place your next answer on top of this one, keeping your answers in order so that the answer to the last test question will be facing you in the ANSWER SPACE.
6. It is very important that you answer every question. Insert a blank card (from STORAGE) if you do not want to guess at an answer, or do not finish the test.
7. Place both envelopes (together) where your teacher tells you. It doesn't matter which is first. DO NOT SEAL THE ENVELOPES.

(Step 6 is ABSOLUTELY required as the computer counts cards and then grades the answers in order. It assumes that the card following Answer #1, is #2.)

THE ENVELOPES CONTAIN:



(The computer reads only the holes in the cards. You may write on the cards to show the question number or other information.)

WHAT TYPE QUESTION CAN BE GRADED BY THE SYSTEM?

The system will grade questions that have any of the following as alternatives:

A B C D

1 2 3 4

True False

Blank (A special card on which none of the above alternatives appear in the answer space. To the computer, the card is "blank.")

SAMPLE QUESTIONS THAT FIT THE SYSTEM

1. Add 3
   3 A. 33 B. 0 C. 6 D. 9

2. Which of the following is spelled incorrectly?
   A. sitting B. spaceing C. running

3. How many of the following are spelled correctly:
   dog, cat, rat, hag
   A. only one B. only two C. only three
   D. all are correct

4. The main character of our story was:
   A. devious B. smart C. sensitive
   D. alert

TRUE/FALSE

____5. The shortest distance between two points is a straight line.

MATCH THE FOLLOWING

6. Lincoln	A. First president
7. Washington	B. War Between the States
8. Truman	C. WWII
	D. Followed Kennedy

HOW DOES THE SYSTEM WORK?

The teacher answer deck is the key to the system. The questions answered by the teacher must also be answered by the student. In effect, the computer looks to see how well the student answer deck matches the teacher answer deck. Therefore, if you do not answer a question, the students must be instructed to ignore the question when compiling a deck.

(For example, let's say that the best form for question ten was a "fill-in-the-blank." You would skip the question when making your deck, and you would instruct the students to skip the question when making their decks.)

For best results, questions to be graded by the computer should be grouped within the test, with the best location being the first part of the test. Yet, it is not necessary that the questions be grouped as the system will grade anything it is told to grade.

The system will grade up to and including fifty questions per student. The questions may be true/false, matching grouped by no more than four choices, multiple-choice, or any type that can be answered as described.

Instructions for loading the decks are located on the envelopes.

THE SYSTEM RATIONALE

Without entering into the argument as to whether the only valid measures of achievement may be obtained through essay testing, fill-in-the-blank, short answer, or other objective testing, we wish to point out some of the attractive features of the objective type test questions that the system is designed to analyze.

First, it may be easily scored. It permits the test maker to cover much more content. And, the performance of all students is judged by the same standard.

It may be argued that guessing distorts the results of objective tests where choices are given. If true, objective tests permit an analysis which will suggest the degree of wild guessing that may have taken place--A test RELIABILITY SCORE. If you can construct a test that yields a good reliability score, you may safely conclude that wild guessing was not a factor that could have distorted the scores of students on your test.

The principle is simple. If a test is composed of sufficiently large numbers of questions that have moderate difficulty for the students, and can discriminate between good and poor students, a highly reliable measure of achievement of students can be obtained.

Then, once a test has been determined to be reliable, you can place more faith in its ability to determine the different levels of learning that have taken place within your class. Improving the reliability of the test, in effect, permits the students who have prepared for the exam to do better than those that have not prepared. You can then better procede to give valid grades to your students.

QUESTIONS????

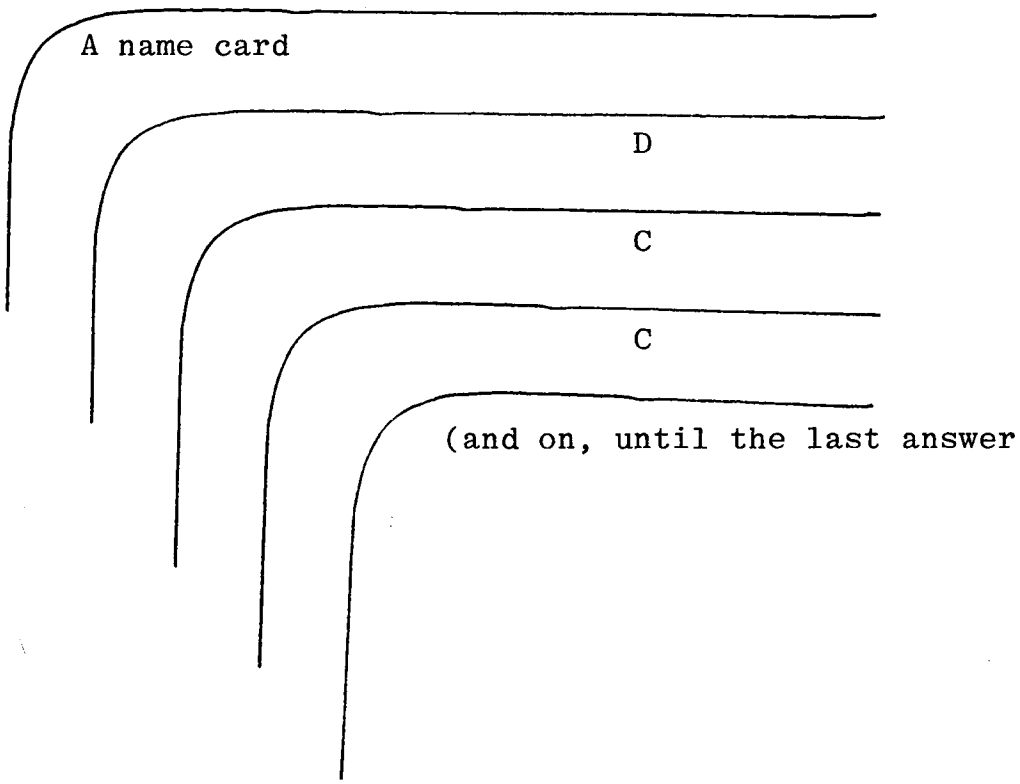
Weekdays
until 3:30* 483-7401 (charge . . . .)

Weekends or
weekdays after 4:00
until 10:00 p.m.* . . . .

*If long distance, dial 0, then the number. Ask the operator to reverse the charges. Place the call person-to-person to Canton Napier if calling 483-7401.

The System will grade the test if:

1. You have loaded a teacher deck.
2. The students have turned in two envelopes--One for the computer, and one for storage.
3. The answer decks look like this:



APPENDIX B

TEST ANALYSIS-FEEDBACK SYSTEM

Thank you for participating in the development of a flexible and inexpensive system that will provide teachers with information about their classroom testing.

You often have to spend hour upon hour of your valuable and limited time scoring tests, either at home or during class when you could be teaching. Although this problem has long been recognized, we do not have much information as to what can be done about it.

But, one thing we do know is that any effort to discover how a system could help with classroom testing has to involve you, and it has to infringe very little on your already limited time.

We believe that together we can discover the techniques and information that need to be a part of any system that is designed to help you with the tedious task of student evaluation.

We need actual student test data to be able to determine what information will be most useful as feedback to teachers. You can help by permitting us to have a copy of your answers (we do not need a copy of the questions), and by permitting us to have a copy of the student answers.

We will return the student answers to you at your convenience, or as suggested.

The data will remain confidential. It will never be identified by teacher or student.

Canton Napier

The University of Tennessee

APPENDIX C

WHAT IS TEST RELIABILITY?

Reliability is the concept that relates to the ability of a test to consistently spread the scores of the individuals who take the test. By consistently, we mean that if other items covering the same area were given, the students who obtained high scores on test one would also obtain high scores on test two. It can be seen that the higher the reliability of the test, the more that test can be a basis for determining student achievement.

Reliability is expressed as a coefficient that may range up to 1.0. It is not to be considered as a percent. Therefore, a reliability of .67 could not be expressed as 67 percent reliable.

Satisfactory reliability for a test is considered to be no less than .90 by test publishers. Therefore, as classroom teachers, we should try to achieve such a high standard, and never settle for any coefficient less than .70.

For a given class of students, reliability may be increased by improving items that are too hard, too easy, or non-discriminating,--concepts that relate to item difficulty and item discrimination. (More on item difficulty and item discrimination below, but first we wish to suggest some general techniques that can help improve reliability.)

HOW MAY TEST RELIABILITY (THE ABILITY OF THE TEST TO CONSISTENTLY SPREAD THE SCORES OF INDIVIDUALS) BE IMPROVED?

The following suggestions may improve your test's reliability:

____ First: Increase the number of questions.

____ Or: Ask more questions about the same objective or topic.

____ And: Work to improve the discrimination of each item; that is, look at past items of yours that DISCRIMINATED well between good and poor students. Write more of the same type for future tests.

____ And: Use items that have a MIDDLE DIFFICULTY rather than items that are too easy or too difficult. That is, start making sure that your tests do not have a floor or ceiling effect (Floor effect = too many hard items; Ceiling effect = too many easy items.)

One warning is necessary concerning test reliability: The wider the range of abilities within a group (more heterogeneous) the higher will be the reliability coefficient for that group's scores. Further, a truly criterion-referenced, and well taught test, may reflect a low coefficient even though the test is reliable.

HOW DO YOU GET "ACCEPTABLE" ITEM DIFFICULTY?

As used within our System, a difficulty score may range from 0 (extremely easy), to 100 (extremely difficult). The optimum difficulty score will be reported to you as 50. When possible, you will want to write items that have a middle difficulty range of from 25 to 75. Contrary to some thinking, tests should not be composed of items that vary in difficulty from very easy to very hard. (Some easy questions may be used at the beginning to a test to put the students at ease.)

To get acceptable item difficulty you may:

_____Offer plausible choices or statements for all alternatives with matching or multiple-choice. That is, don't use ideas that are too obviously wrong.

_____Concern yourself with writing quality choices for answers, rather than a quantity of choices.

_____When using True/False, write an equal number of each.

_____If the item has a "stem" (as in the beginning of a multiple-choice question), see that it is not too complex nor simple.

_____An item can be made more difficult, if multiple-choice, by making alternatives more homogeneous.

Items of acceptable difficulty have the potential for discriminating between the students who studied for the test and those who didn't. A TEST FULL OF ITEMS THAT DISCRIMINATE PERMITS THE TEST TO DISCRIMINATE.

HOW DO WE DETERMINE ITEM DISCRIMINATION?

Item discrimination works this way. To determine if an item is contributing to the RELIABILITY of your test, our System looks at those students who did the best and the poorest on a test, then selects the top and bottom 27 percent. If all those who made a high score got a question correct, and all those who made a low test score missed it, the DISCRIMINATION will be reported as one (1.0), or "perfect discrimination." Conversely, if students with poor scores get a question, and those with good scores miss it, the discrimination will be minus one (-1.0), or less than perfect.

Discrimination scores may be considered in four categories:

HIGH	if the score is .40 or higher
MODERATE	if the score is .20 to .39
LOW	if the score is .01 to .19
POOR	if the score is .00 or negative

The minimum desirable score is considered to be thirty (+.30).

Many of your students will be interested in this information about their test, especially as to which questions were (according to the computer) "too hard," or "too easy." You may wish to use the data to stimulate a class review of the test. This information, or the ITEM EASE score and graph below, can be used. You may be pleasantly surprised to learn that the students have a feel for questions that contributed least to the validity of the test.

APPENDIX D

****

UNIVERSITY OF TENNESSEE

CANTON NAPIER
SIMULATION

STUDENT	ID#	PERCENT CORRECT	NUMBER CORRECT	ITEM ANALYSIS (1=CORRECT ANSWER 0=INCORRECT ANSWER)
LISA A	(101)	83%	5	11110 1
LISA B	(102)	33%	2	00101 0
LISA C	(103)	33%	2	00010 1
LISA D	(104)	50%	3	00111 0
LISA E	(105)	67%	4	11110 0
LISA F	(106)	50%	3	01101 0
LISA G	(107)	17%	1	01000 0
LISA H	(108)	100%	6	11111 1
LISA I	(109)	50%	3	01101 0
LISA J	(110)	67%	4	11100 1

****

UNIVERSITY OF TENNESSEE

CANTON NAPIER
SIMULATION

CLASS AVERAGE	TOTAL OF PERCENT CORRECT FOR TEST	STUDENTS IN CLASS	AVERAGE
55.00	550.00	10	55.00
55.00	550.00	10	55.00
55.00	550.00	10	55.00
55.00	550.00	10	55.00
55.00	550.00	10	55.00
55.00	550.00	10	55.00
55.00	550.00	10	55.00
55.00	550.00	10	55.00
55.00	550.00	10	55.00
55.00	550.00	10	55.00
55.00	550.00	10	55.00
55.00	550.00	10	55.00
55.00	550.00	10	55.00
55.00	550.00	10	55.00
55.00	550.00	10	55.00
55.00	550.00	10	55.00
55.00	550.00	10	55.00
55.00	550.00	10	55.00
55.00	550.00	10	55.00
55.00	550.00	10	55.00

****

UNIVERSITY OF TENNESSEE

CANTON NAPIER
SIMULATION

RELIABILITY 0.47
RELIABILITY 0.47
RELIABILITY 0.47
RELIABILITY 0.47
RELIABILITY 0.47
RELIABILITY 0.47
RELIABILITY 0.47
RELIABILITY 0.47
RELIABILITY 0.47
RELIABILITY 0.47

RELIABILITY IS THE CONCEPT THAT RELATES TO THE ABILITY OF A TEST TO CONSISTENTLY SPREAD THE SCORES OF THE INDIVIDUALS WHO TAKE THE TEST. BY CONSISTENTLY, WE MEAN THAT IF OTHER ITEMS COVERNING THE SAME AREA WERE GIVEN, THE STUDENTS WHO OBTAINED HIGH SCORES CN TEST CNE WOULD ALSO OBTAIN HIGH SCORES ON TEST TWO. IT CAN BE SEEN THAT THE HIGHER THE RELIABILITY OF THE TEST, THE MORE THAT TEST CAN BE A BASIS FOR DETERMINING STUDENT ACHIEVEMENT.

RELIABILITY IS EXPRESSED AS A COEFFICIENT THAT MAY RANGE UP TO 1.0. IT IS NOT TO BE CONSIDERED AS A PERCENT. THEREFORE, A RELIABILITY OF .67 COULD NOT BE EXPRESSED AS 67% RELIABLE.

SATISFACTORY RELIABILITY FOR A TEST IS CONSIDERED TO BE NO LESS THAN .90 BY TEST PUBLISHERS. THEREFORE, AS CLASSROOM TEACHERS, WE SHOULD TRY TO ACHIEVE SUCH A HIGH STANDARD, AND NEVER SETTLE FOR ANY COEFFICIENT LESS THAN .70.

FOR A GIVEN CLASS OF STUDENTS, RELIABILITY MAY BE INCREASED BY IMPROVING ITEMS THAT ARE TOO HARD, TOO EASY, OR NON-DISCRIMINATING,-- CONCEPTS THAT RELATE TO ITEM DIFFICULTY AND ITEM DISCRIMINATION.

RELIABILITY 0.47
RELIABILITY 0.47
RELIABILITY 0.47
RELIABILITY 0.47
RELIABILITY 0.47
RELIABILITY 0.47
RELIABILITY 0.47
RELIABILITY 0.47
RELIABILITY 0.47
RELIABILITY 0.47

THE FOLLOWING SUGGESTIONS MAY IMPROVE YOUR TEST'S RELIABILITY:

FIRST: INCREASE THE NUMBER OF QUESTIONS.

OR: ASK MORE QUESTIONS ABOUT THE SAME OBJECTIVE OR TOPIC.

AND: WORK TO IMPROVE THE DISCRIMINATION OF EACH ITEM; THAT IS, LOOK AT PAST ITEMS OF YOURS THAT DISCRIMINATED WELL BETWEEN GOOD AND POOR STUDENTS. WRITE MORE OF THE SAME TYPE FOR FUTURE TESTS.

AND: USE ITEMS THAT HAVE A MIDDLE DIFFICULTY RATHER THAN ITEMS THAT ARE TOO EASY OR TOO DIFFICULT. THAT IS, START MAKING SURE THAT YOUR TESTS DO NOT HAVE A FLOOR OR CEILING EFFECT (FLOOR EFFECT= TOO MANY HARD ITEMS; CEILING EFFECT= TOO MANY EASY ITEMS.)

ONE WARNING IS NECESSARY CONCERNING TEST RELIABILITY: THE WIDER THE RANGE OF ABILITIES WITHIN A GROUP (MORE HETEROGENEOUS) THE HIGHER WILL BE THE RELIABILITY COEFFICIENT FOR THAT GROUP'S SCORES. FURTHER, A TRULY CRITERION-REFERENCED, AND WELL TAUGHT TEST, MAY REFLECT A LOW COEFFICIENT EVEN THOUGH THE TEST IS RELIABLE.

****

UNIVERSITY OF TENNESSEE

CANTON NAPIER
SIMULATION

MIDDLE DIFFICULTY RANGE
****
****
****
****
****
****

DESIRABLE DISCRIMINATION RANGE
****
****
****
****
****
****

ITEM	1)	DIFFICULTY 50.0		(ITEM	1)	DISCRIMINATION 1.00
		**********				**********
ITEM	2)	DIFFICULTY 33.3		(ITEM	2)	DISCRIMINATION 0.67
		**********				**********
ITEM	3)	DIFFICULTY 33.3		(ITEM	3)	DISCRIMINATION 0.67
		**********				**********
ITEM	4)	DIFFICULTY 33.3		(ITEM	4)	DISCRIMINATION 0.67
		**********				**********
ITEM	5)	DIFFICULTY 66.7		(ITEM	5)	DISCRIMINATION 0.00
		**********				X
ITEM	6)	DIFFICULTY 50.0		(ITEM	6)	DISCRIMINATION 0.33
		**********				**********

AS USED WITHIN OUR SYSTEM, A DIFFICULTY SCORE MAY RANGE FROM 0 (EXTREMELY EASY), TO 100 (EXTREMELY DIFFICULT). THE OPTIMUM DIFFICULTY SCORE WILL BE REPORTED TO YOU AS 50. WHEN POSSIBLE, YOU WILL WANT TO WRITE ITEMS THAT HAVE A MIDDLE DIFFICULTY RANGE OF FROM 25 TO 75. CONTRARY TO SOME THINKING, TESTS SHOULD NOT BE COMPOSED OF ITEMS THAT VARY IN DIFFICULTY FROM VERY EASY TO VERY HARD. (SOME EASY QUESTIONS MAY BE USED AT THE BEGINNING OF A TEST TO PUT THE STUDENTS AT EASE.)

TO GET ACCEPTABLE ITEM DIFFICULTY YOU MAY:

OFFER PLAUSIBLE CHOICES OR STATEMENTS FOR ALL ALTERNATIVES WITH MATCHING OR MULTIPLE-CHOICE. THAT IS, DON'T USE IDEAS THAT ARE TOO OBVIOUSLY WRONG.

CONCERN YOURSELF WITH WRITING QUALITY CHOICES FOR ANSWERS, RATHER THAN A QUANTITY OF CHOICES.

WHEN USING TRUE/FALSE, WRITE AN EQUAL NUMBER OF EACH.

IF THE ITEM HAS A "STEM" (AS IN THE BEGINNING OF A MULTIPLE-CHOICE QUESTION), SEE THAT IT IS NOT TOO COMPLEX NOR SIMPLE.

AN ITEM CAN BE MADE MORE DIFFICULT, IF MULTIPLE-CHOICE, BY MAKING ALTERNATIVES MORE HOMOGENEOUS.

ITEMS OF ACCEPTABLE DIFFICULTY HAVE THE POTENTIAL FOR DISCRIMINATING BETWEEN THE STUDENTS WHO STUDIED FOR THE TEST AND THOSE WHO DIDN'T. A TEST FULL OF ITEMS THAT DISCRIMINATE PERMITS THE TEST TO DISCRIMINATE.

ITEM DISCRIMINATION WORKS THIS WAY. TO DETERMINE IF AN ITEM IS CONTRIBUTING TO THE RELIABILITY OF YOUR TEST, OUR SYSTEM LOOKS AT THOSE STUDENTS WHO DID THE BEST AND THE POOREST ON A TEST, THEN SELECTS THE TOP AND BOTTOM 27%. IF ALL THOSE WHO MADE A HIGH SCORE GOT A QUESTION CORRECT, AND ALL THOSE WHO MADE A LOW TEST SCORE MISSED IT, THE DISCRIMINATION WILL BE REPORTED AS ONE (1.0), OR "PERFECT DISCRIMINATION." CONVERSELY, IF STUDENTS WITH POOR SCORES GET A QUESTION, AND THOSE WITH GOOD SCORES MISS IT, THE DISCRIMINATION WILL BE MINUS ONE (-1.0), OR LESS THAN PERFECT.

DISCRIMINATION SCORES MAY BE CONSIDERED IN FOUR CATEGORIES:

HIGH	IF THE SCORE IS .40 OR HIGHER
MODERATE	IF THE SCORE IS .20 TO .39
LOW	IF THE SCORE IS .01 TO .19
POOR	IF THE SCORE IS .00 OR NEGATIVE

THE MINIMUM DESIRABLE SCORE IS CONSIDERED TO BE THIRTY (+.30).

MANY OF YOUR STUDENTS WILL BE INTERESTED IN THE INFORMATION ABOUT THEIR TEST, ESPECIALLY AS TO WHICH QUESTIONS WERE (ACCORDING TO THE COMPUTER) "TOO HARD", OR "TOO EASY." YOU MAY WISH TO USE THE DATA TO STIMULATE A CLASS REVIEW OF THE TEST. THIS INFORMATION, OR THE ITEM EASE SCORE AND GRAPH BELOW, CAN BE USED. YOU MAY BE PLEASANTLY SURPRISED TO LEARN THAT THE STUDENTS HAVE A FEEL FOR QUESTIONS THAT CONTRIBUTED LEAST TO THE VALIDITY OF THE TEST.

UNIVERSITY OF TENNESSEE

CANTON NAPIER
SIMULATION

ITEM	RAW SCORE	PERCENT	++++ITEM EASE++++ (PERCENT OF STUDENTS ANSWERING ITEM CORRECTLY)
(1)	4	40.00	**********
(2)	7	70.00	**********
(3)	8	80.00	**********
(4)	5	50.00	**********
(5)	5	50.00	**********
(6)	4	40.00	**********

****

UNIVERSITY OF TENNESSEE

CANTON NAPIER
SIMULATION

NUMBER CHOOSING EACH ANSWER (*=CORRECT ANSWER)

		FIRST	SECOND	THIRD	FOURTH	FIFTH	SIXTH
ITEM	1	2	4 *	2	2	0	0
ITEM	2	0	2	7 *	1	0	0
ITEM	3	1	8 *	1	0	0	0
ITEM	4	5 *	1	2	2	0	0
ITEM	5	1	2	2	5 *	0	0
ITEM	6	3	2	1	4 *	0	0

APPENDIX E

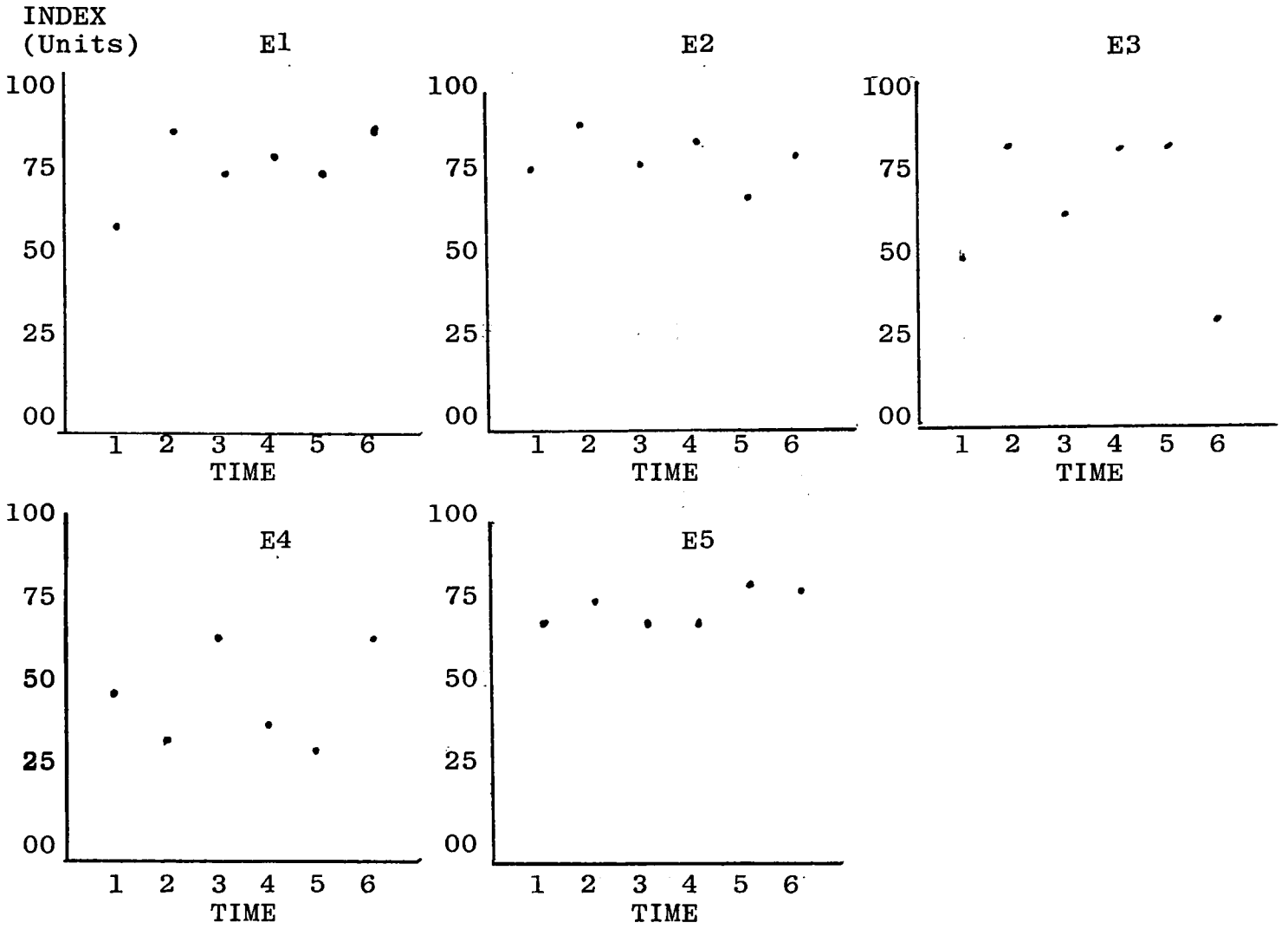


FIGURE 1. Reliability Indices for Ten Subjects (E1 to C10).

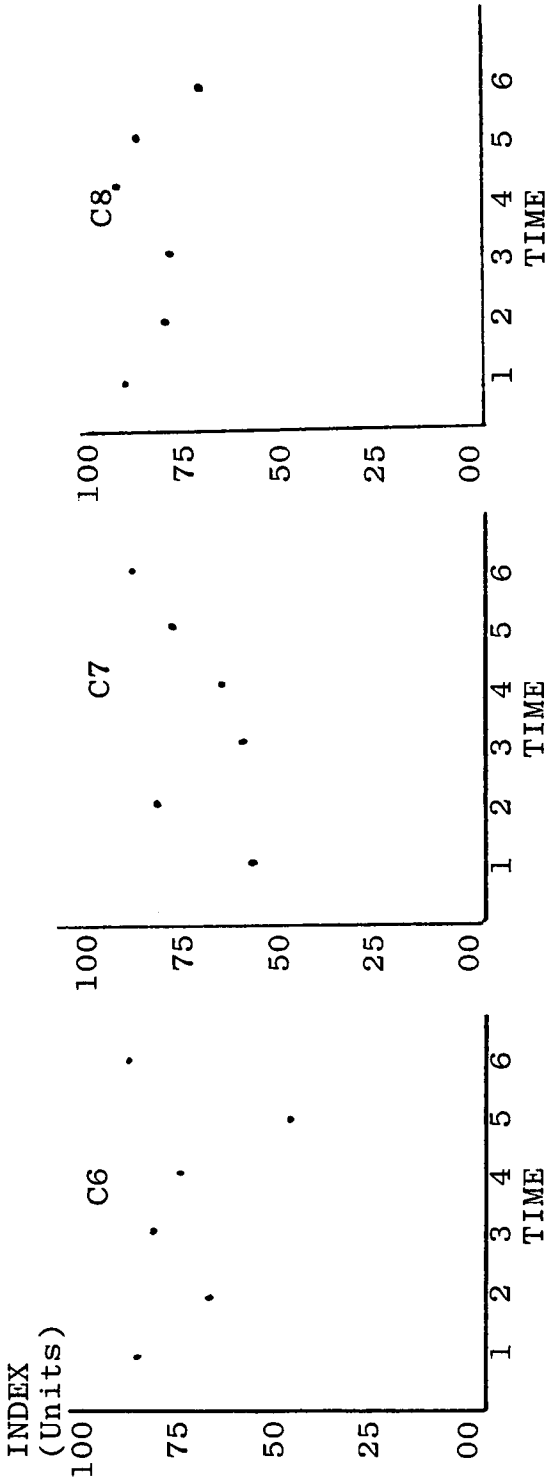


FIGURE 1. (Continued).

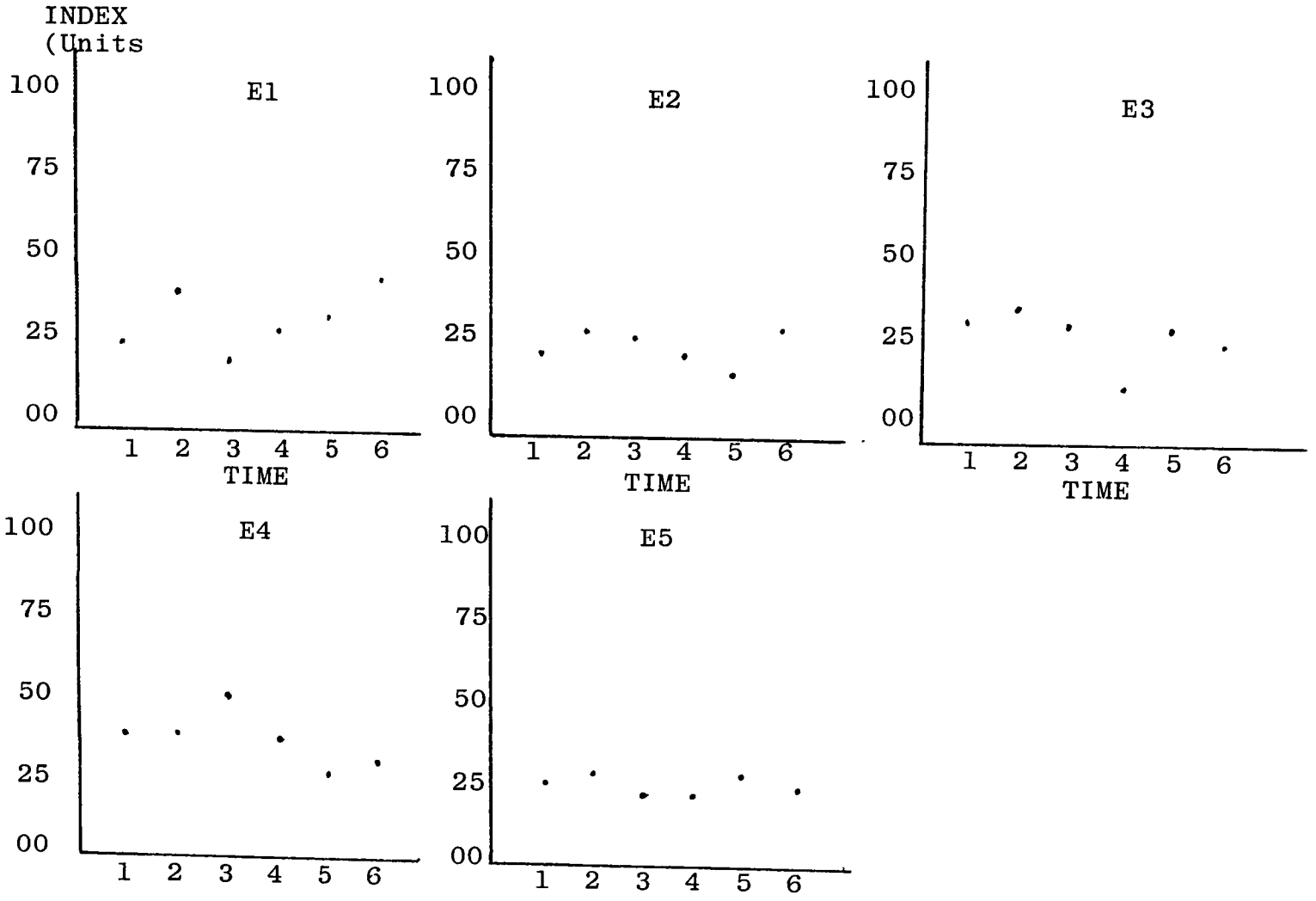


FIGURE 2. Difficulty Indices for Ten Subjects (E1 to C10).

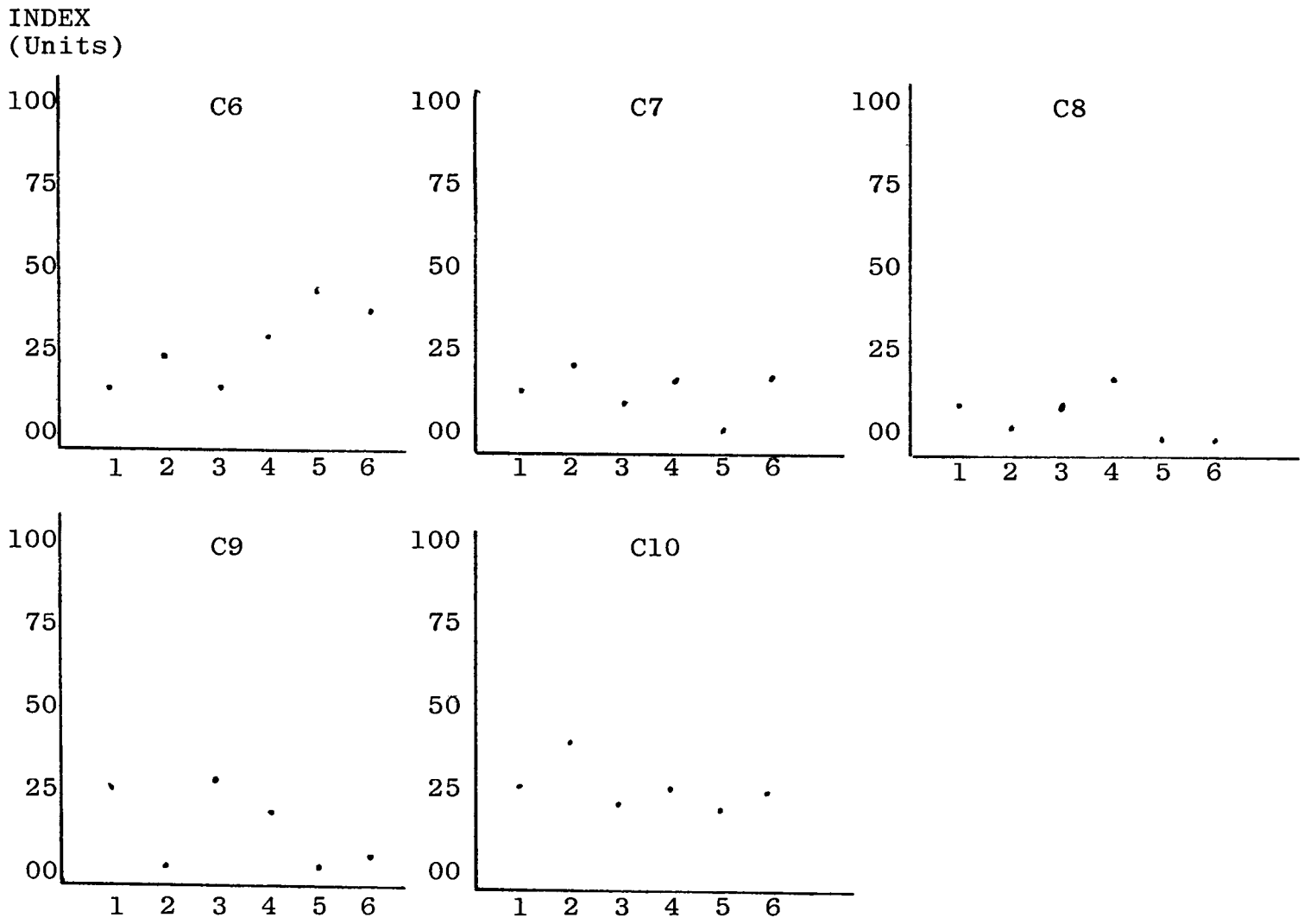


FIGURE 2. (Continued).

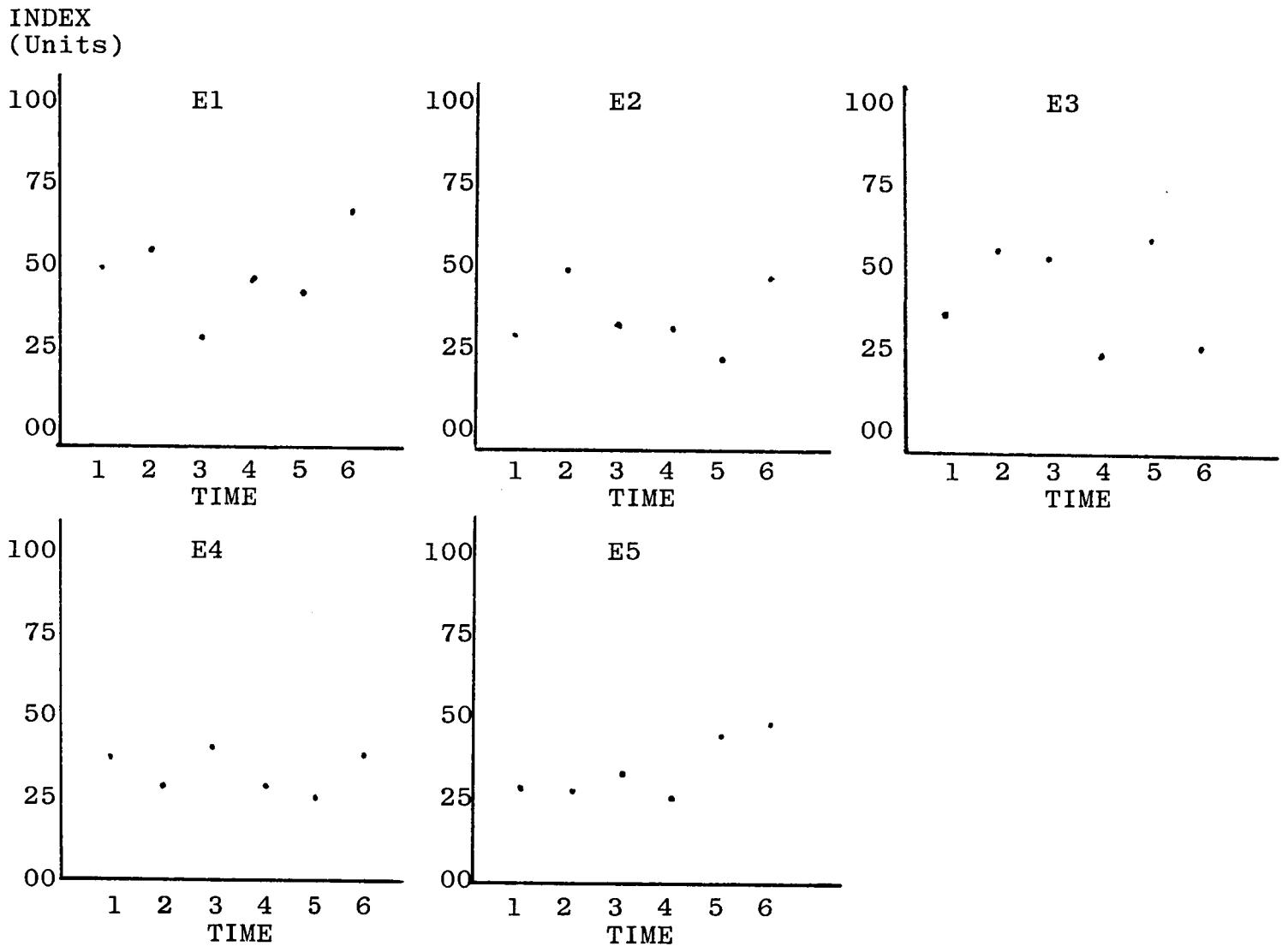


FIGURE 3. Discrimination Indices for Ten Subjects (E1 to C10).

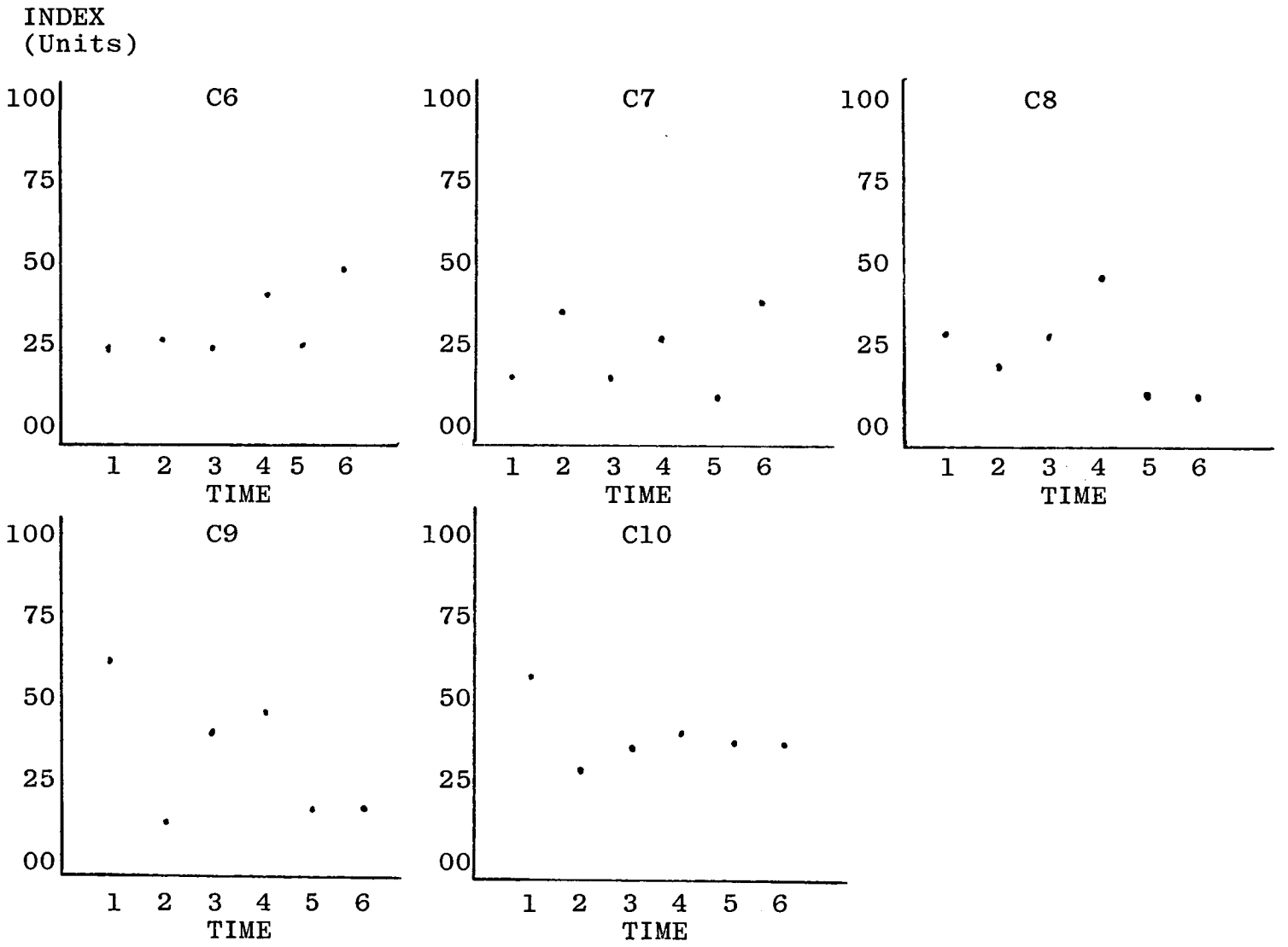


FIGURE 3. (Continued)

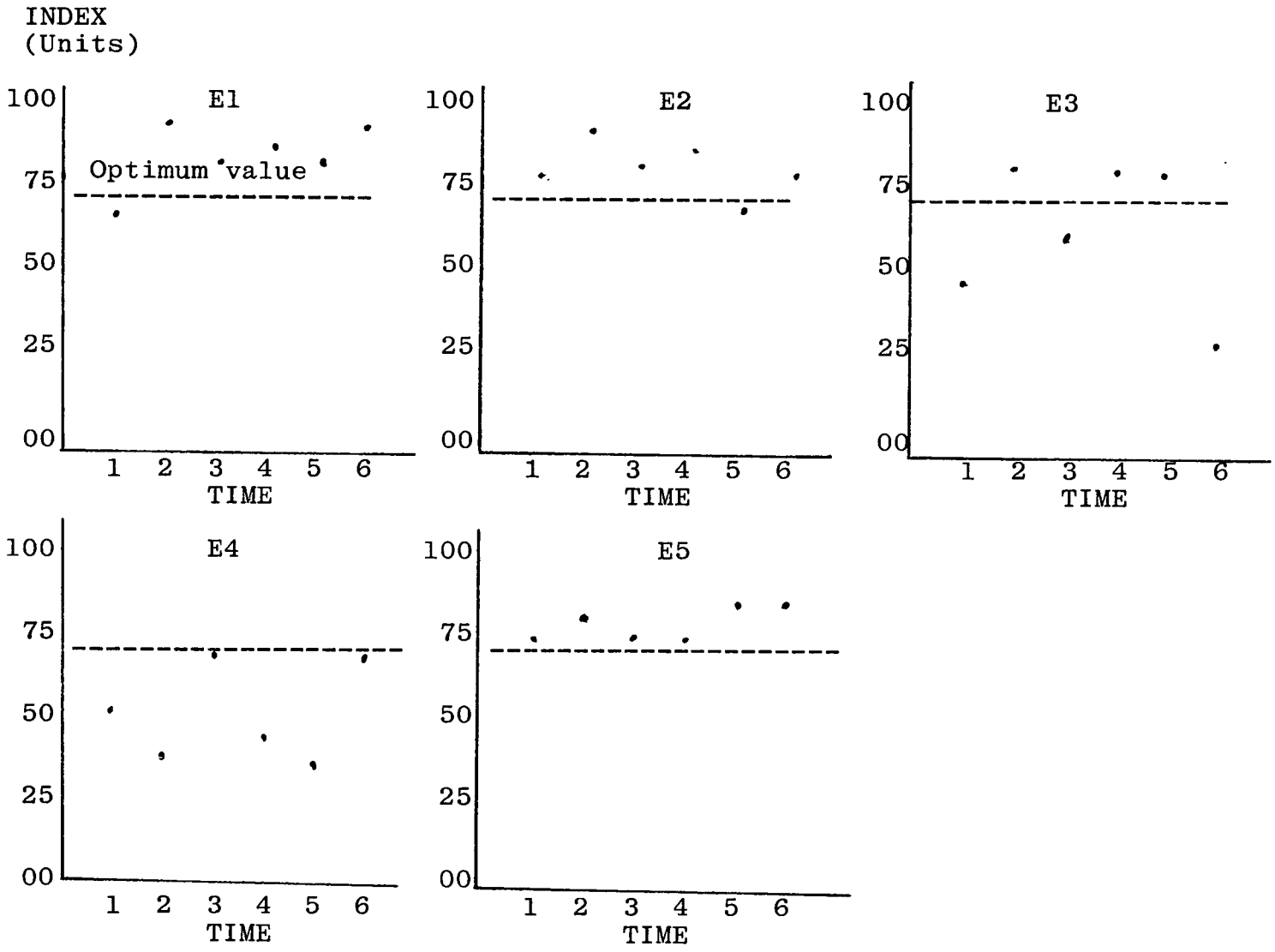


FIGURE 4. Reliability Indices With Optimum Value for Ten Subjects (E1 to C10).

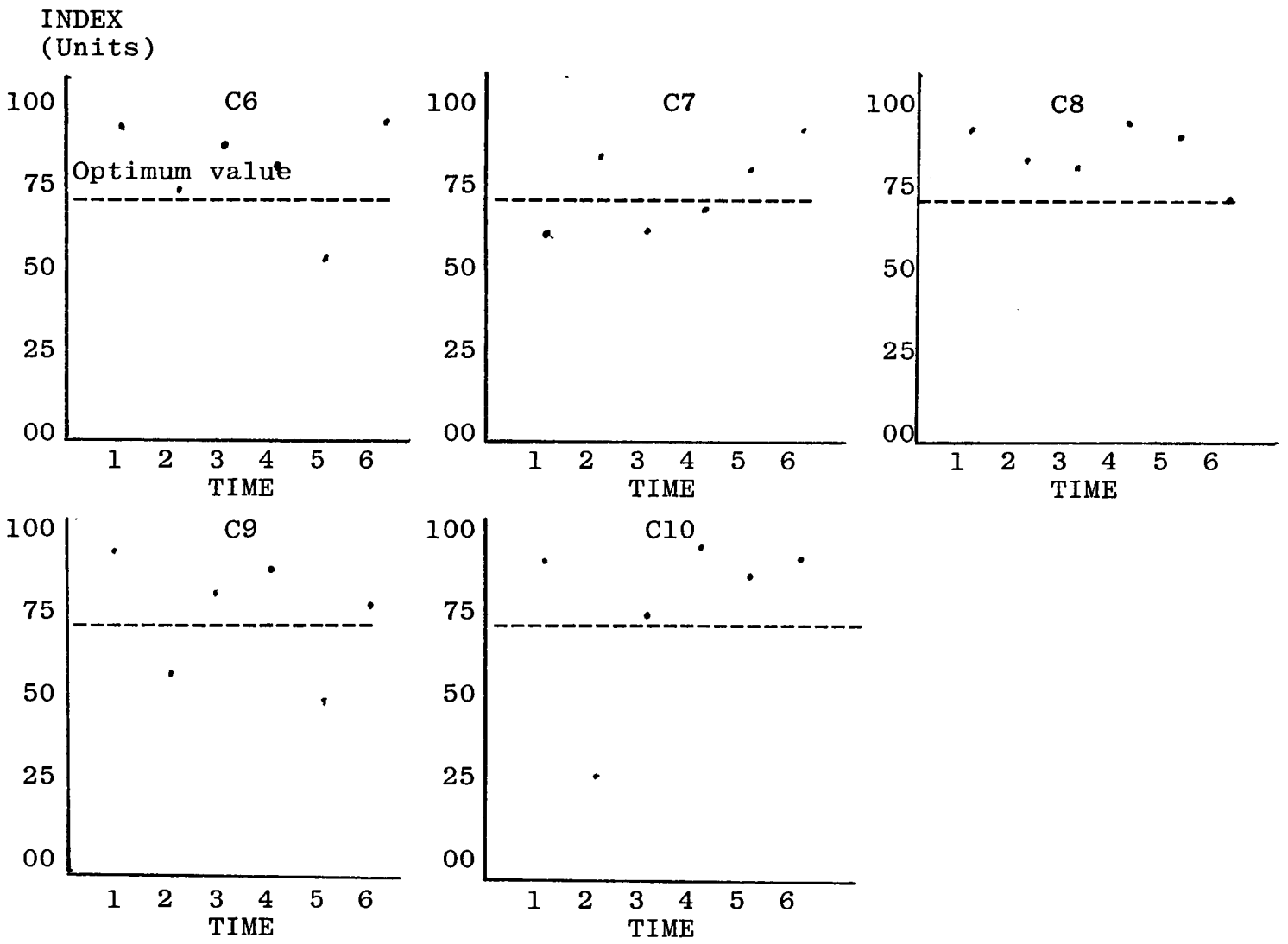


FIGURE 4. (Continued).

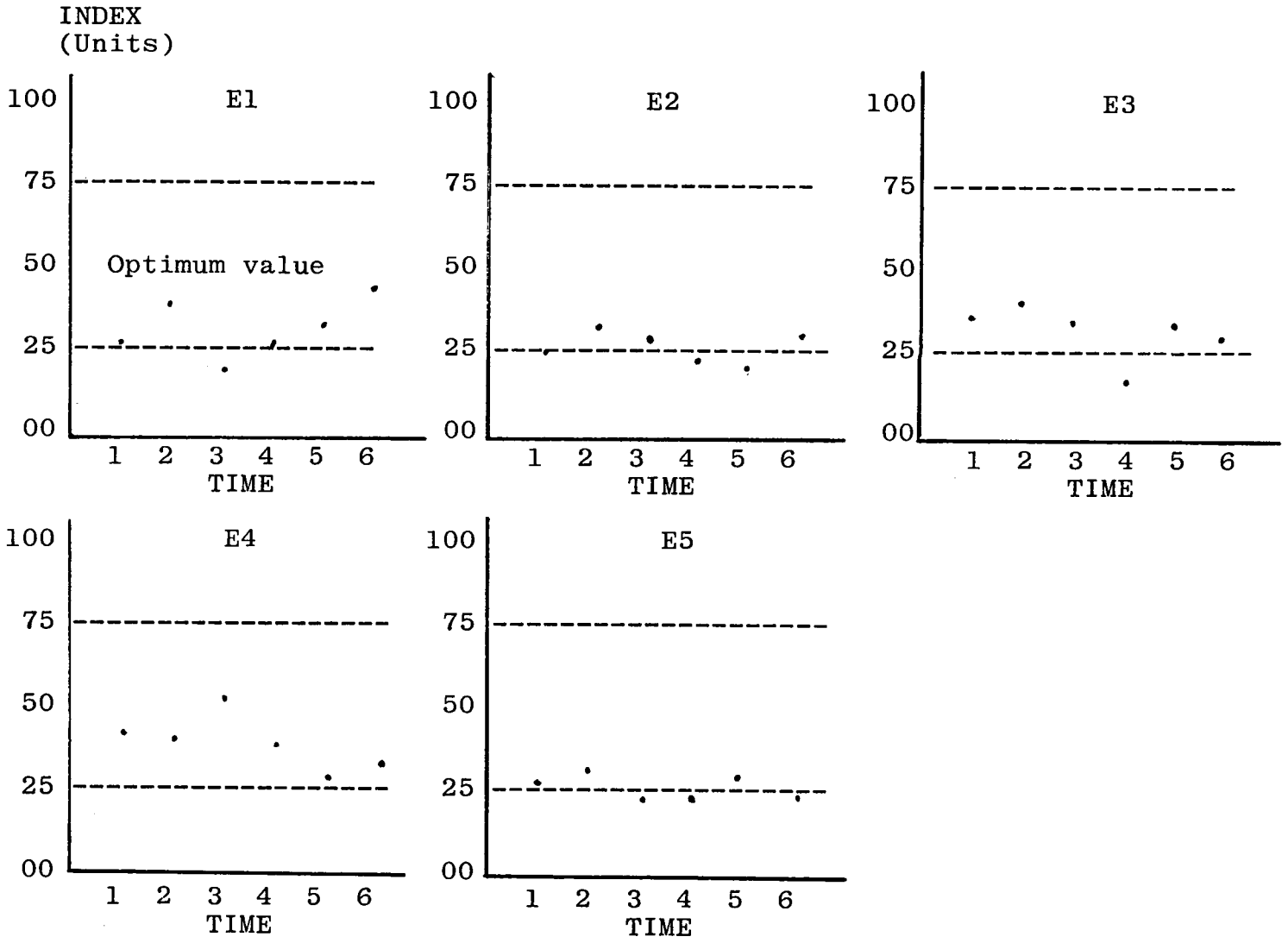


FIGURE 5. Difficulty Indices With Optimum Value for Ten Subjects (E1 to C10).

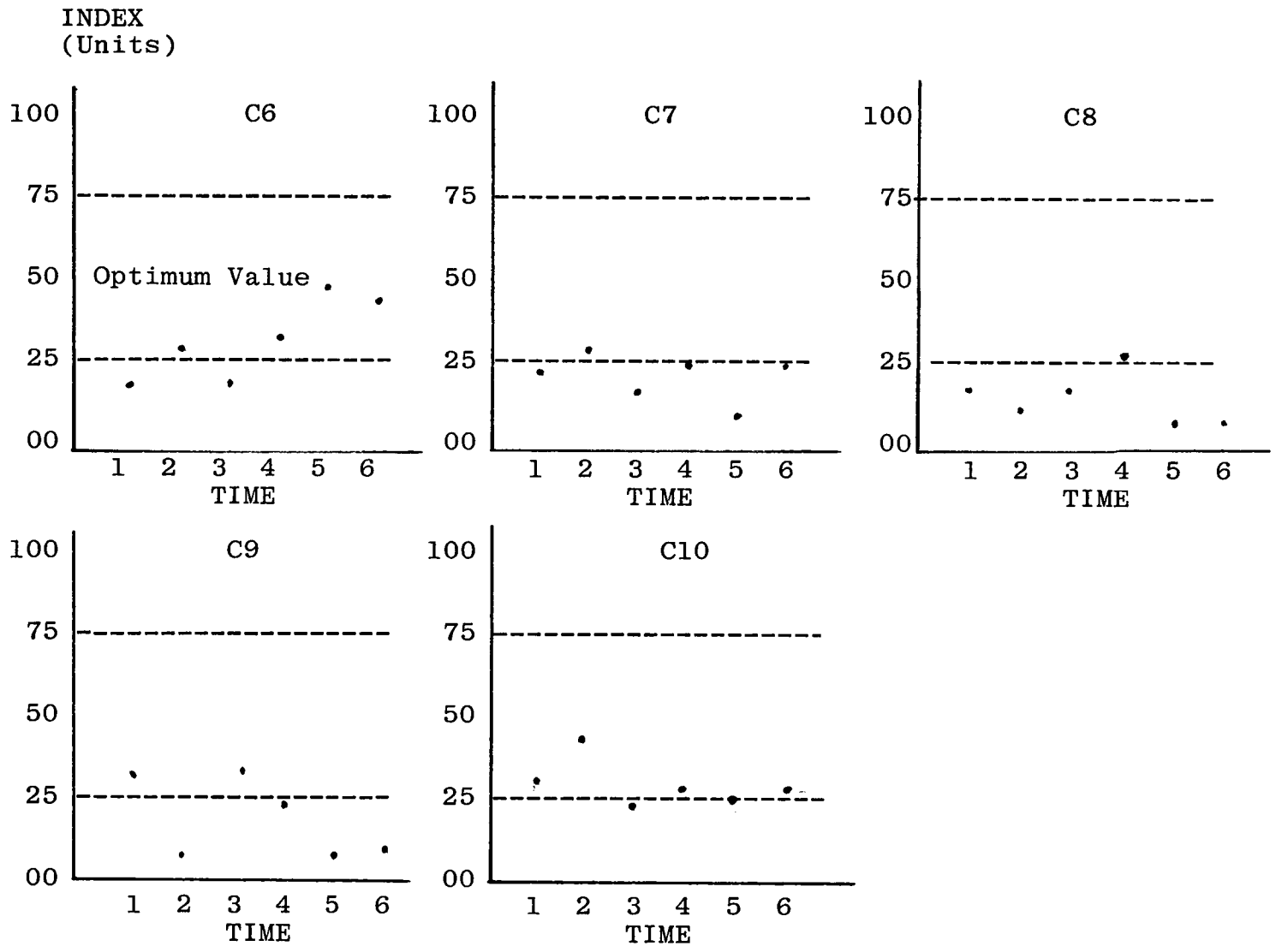


FIGURE 5. (Continued).

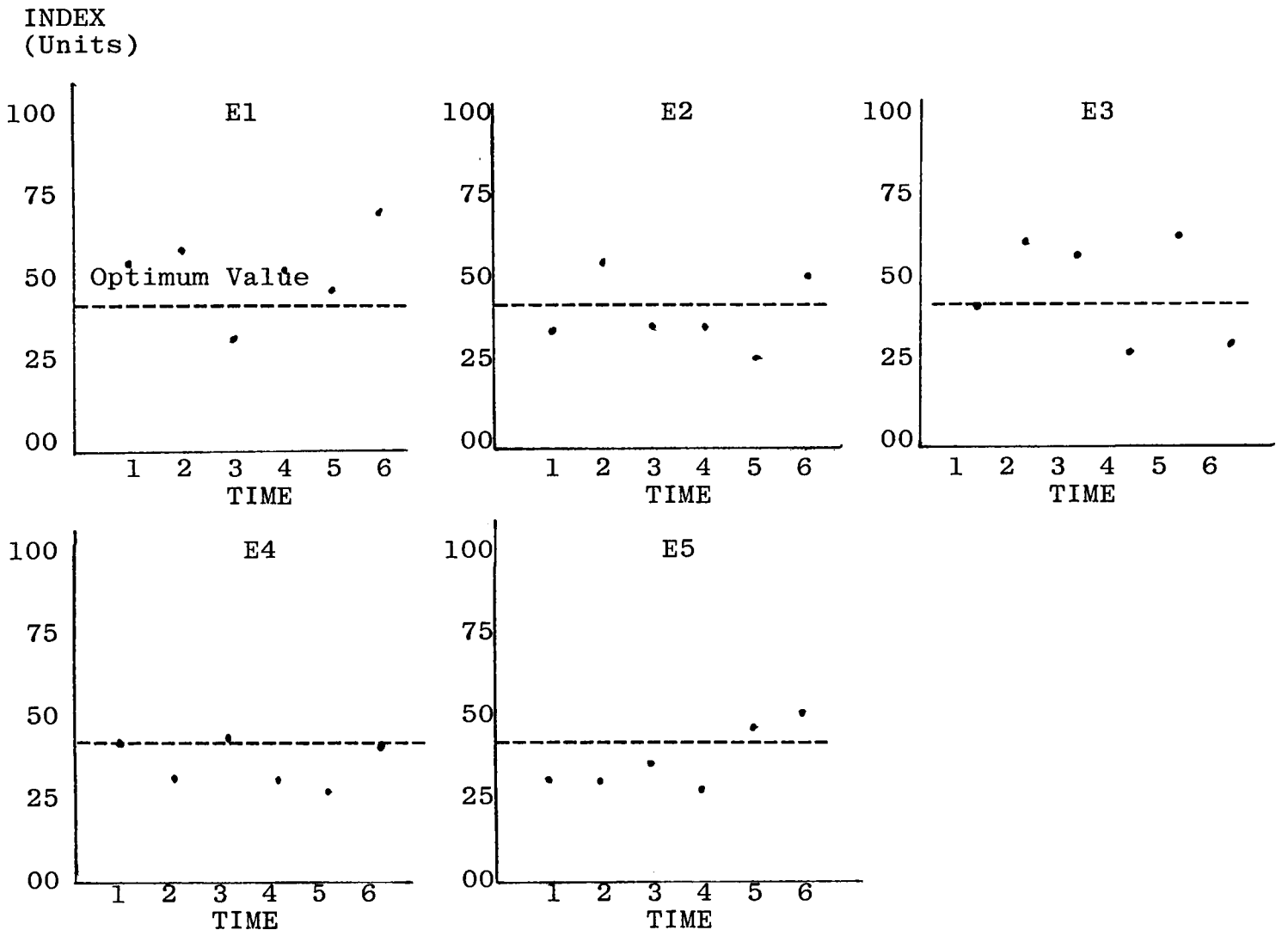


FIGURE 6. Discrimination Indices With Optimum Value for Ten Subjects (E1 to C10).

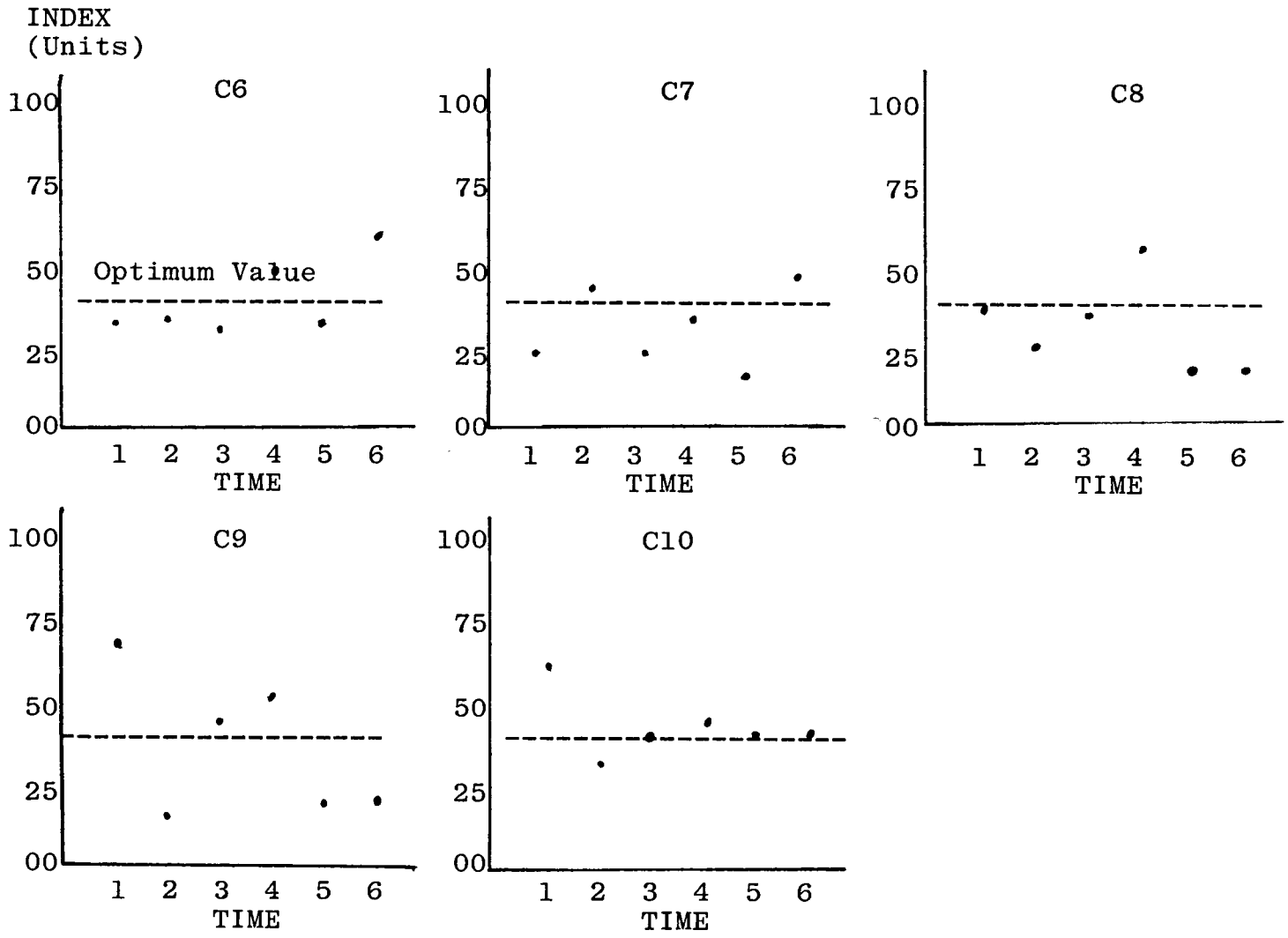


FIGURE 6. (Continued).

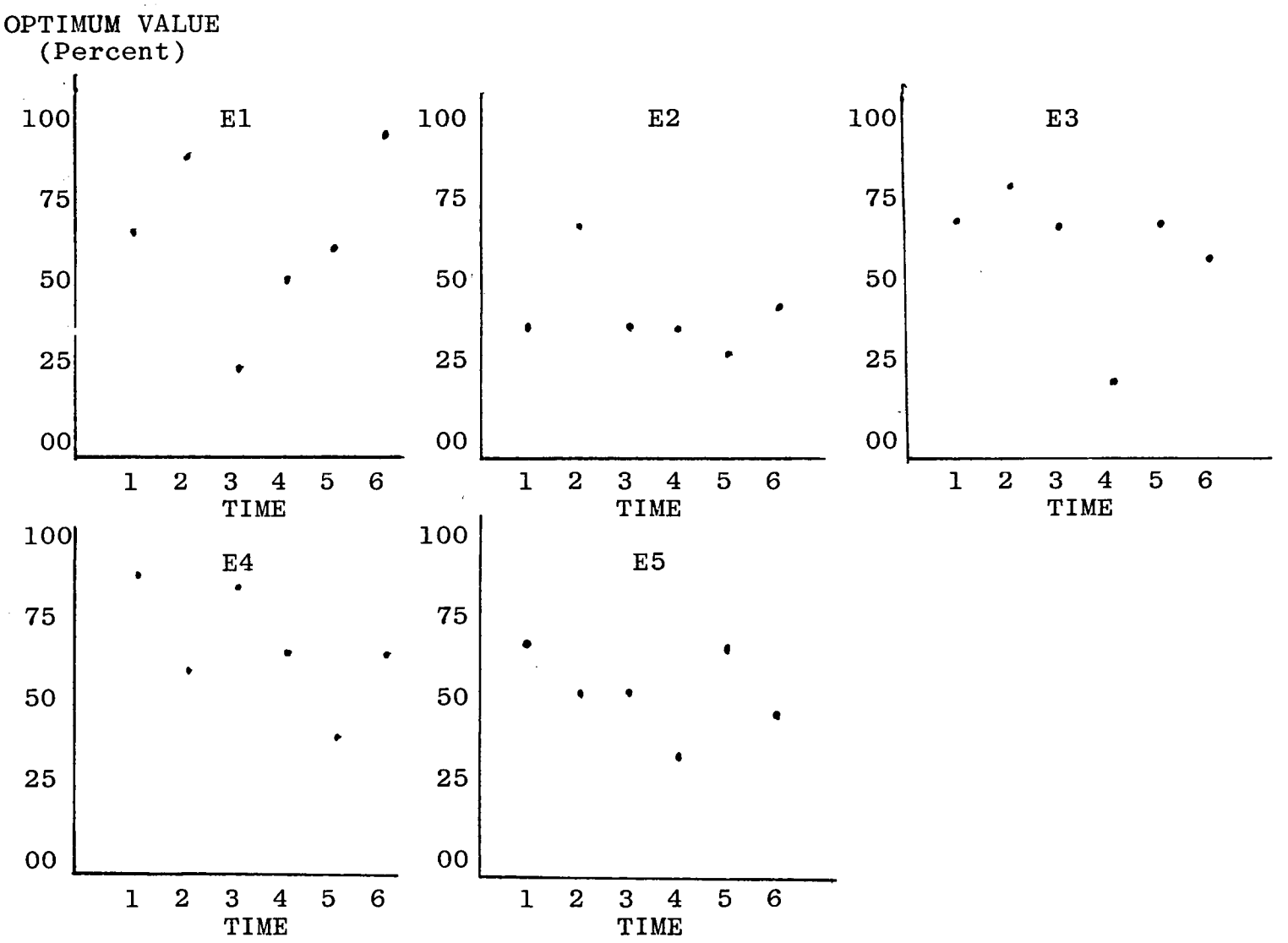


FIGURE 7. Optimum Item Difficulty Achieved by Ten Subjects (E1 to C10).

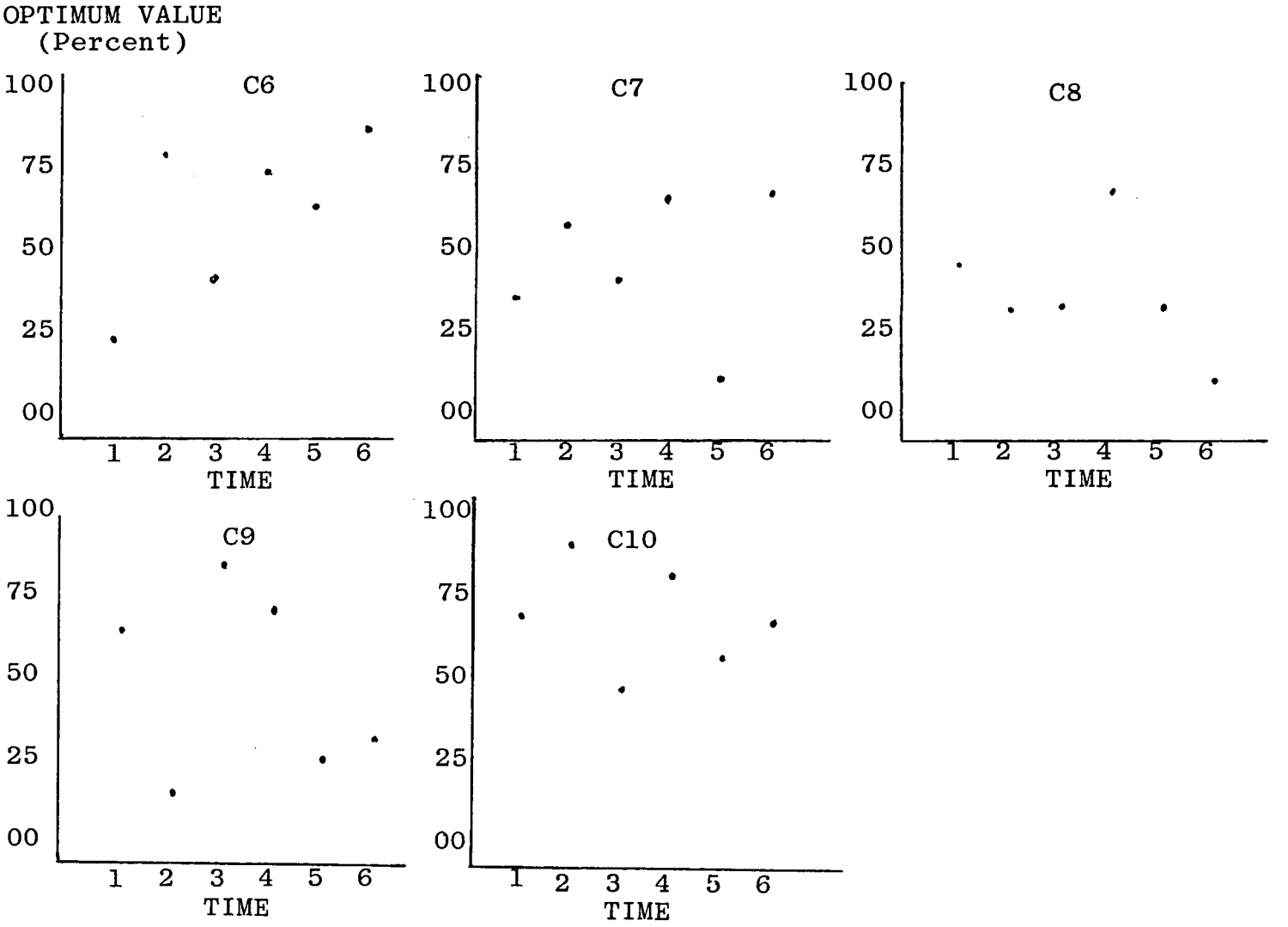


FIGURE 7. (Continued).

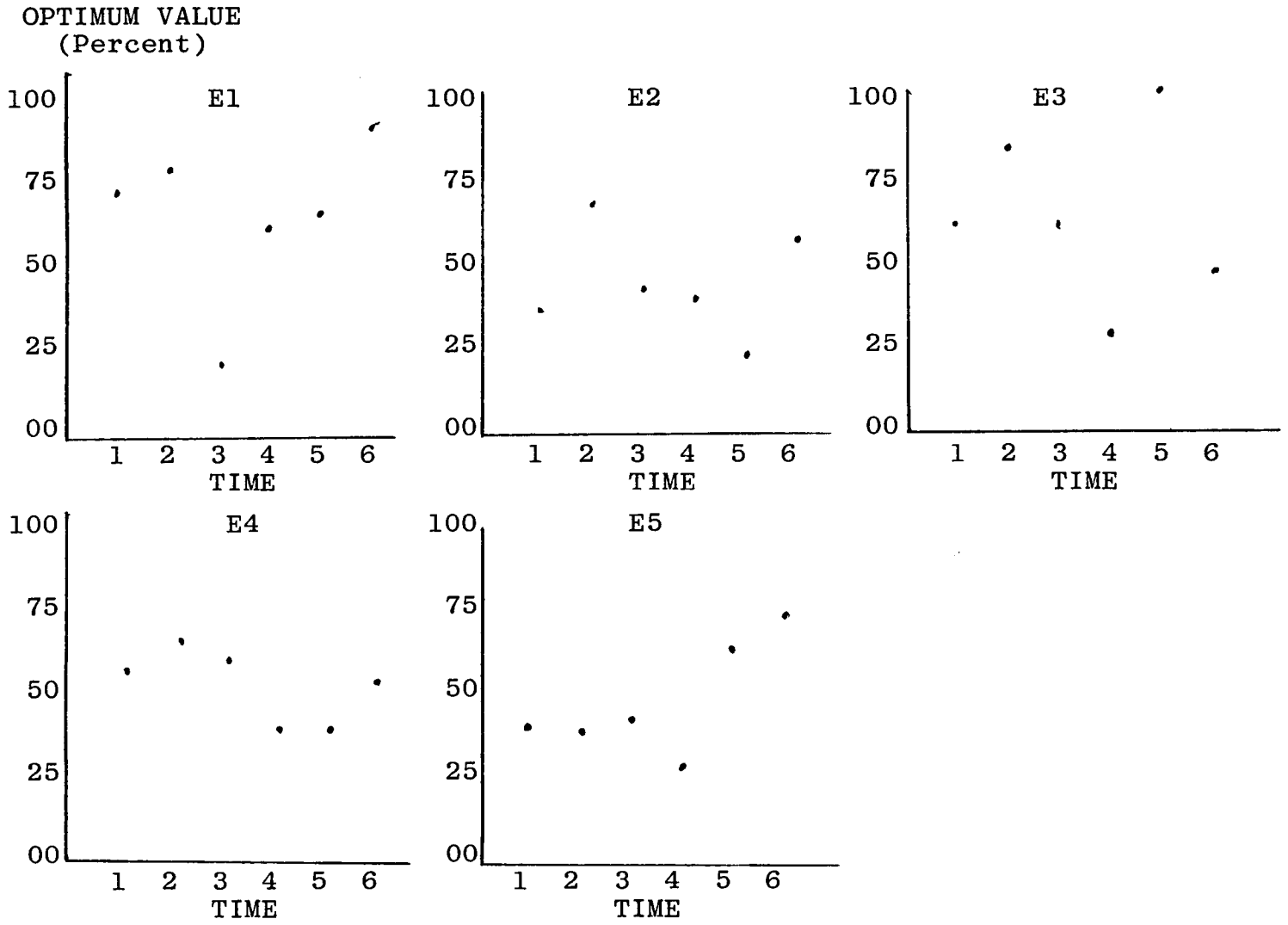


FIGURE 8. Optimum Item Discrimination Achieved by Ten Subjects (E1 to C10).

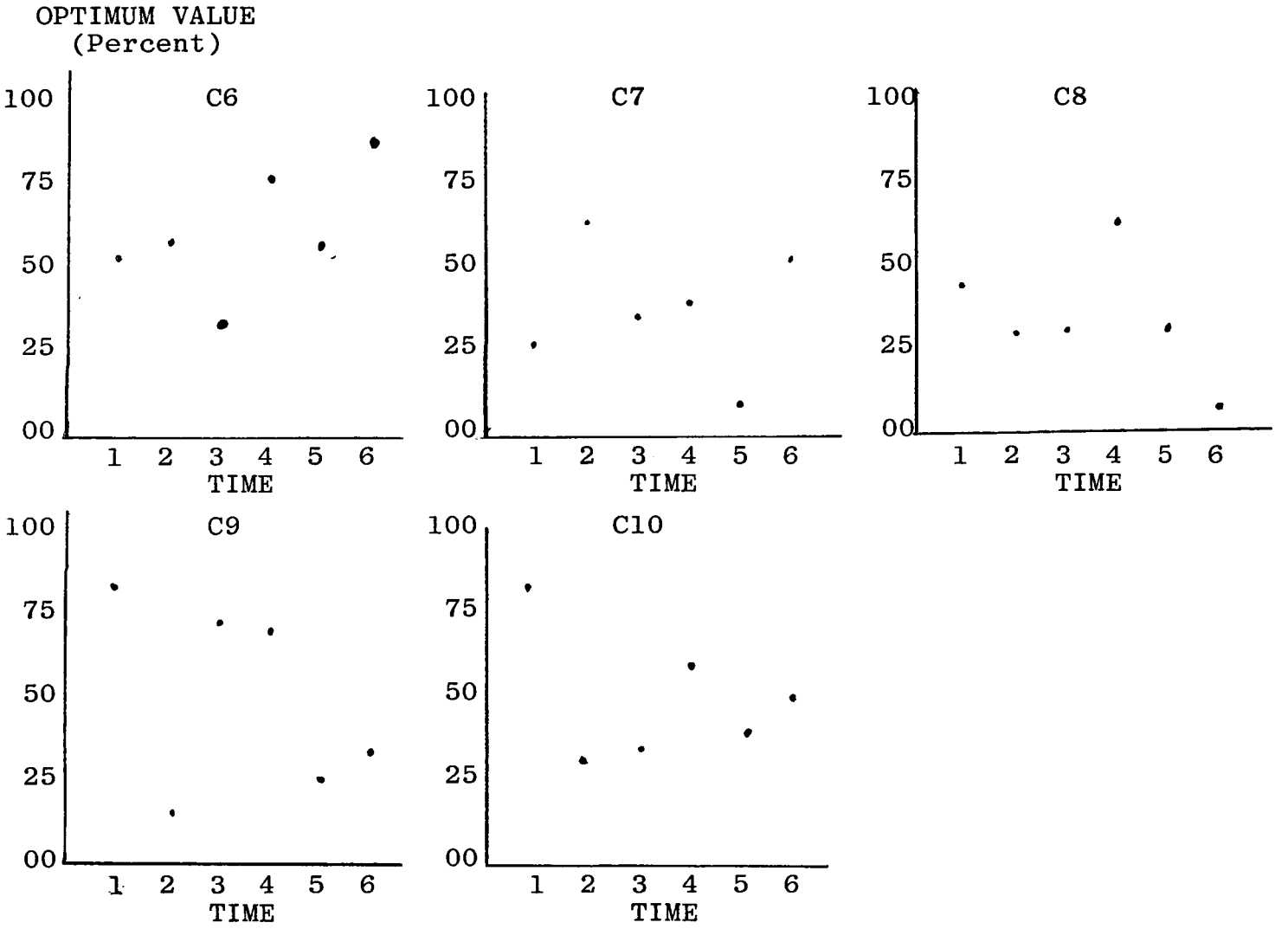


FIGURE 8. (Continued).

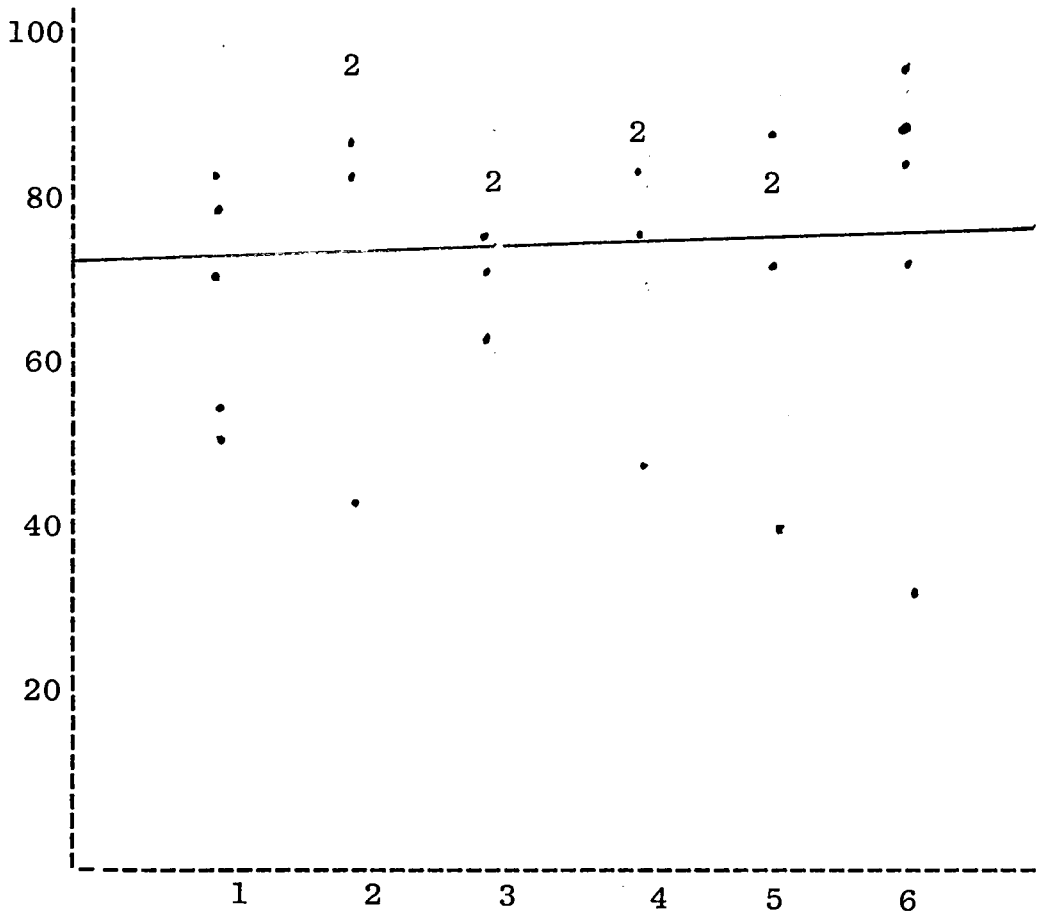


FIGURE 9. Test Reliability for the Experimental Group.

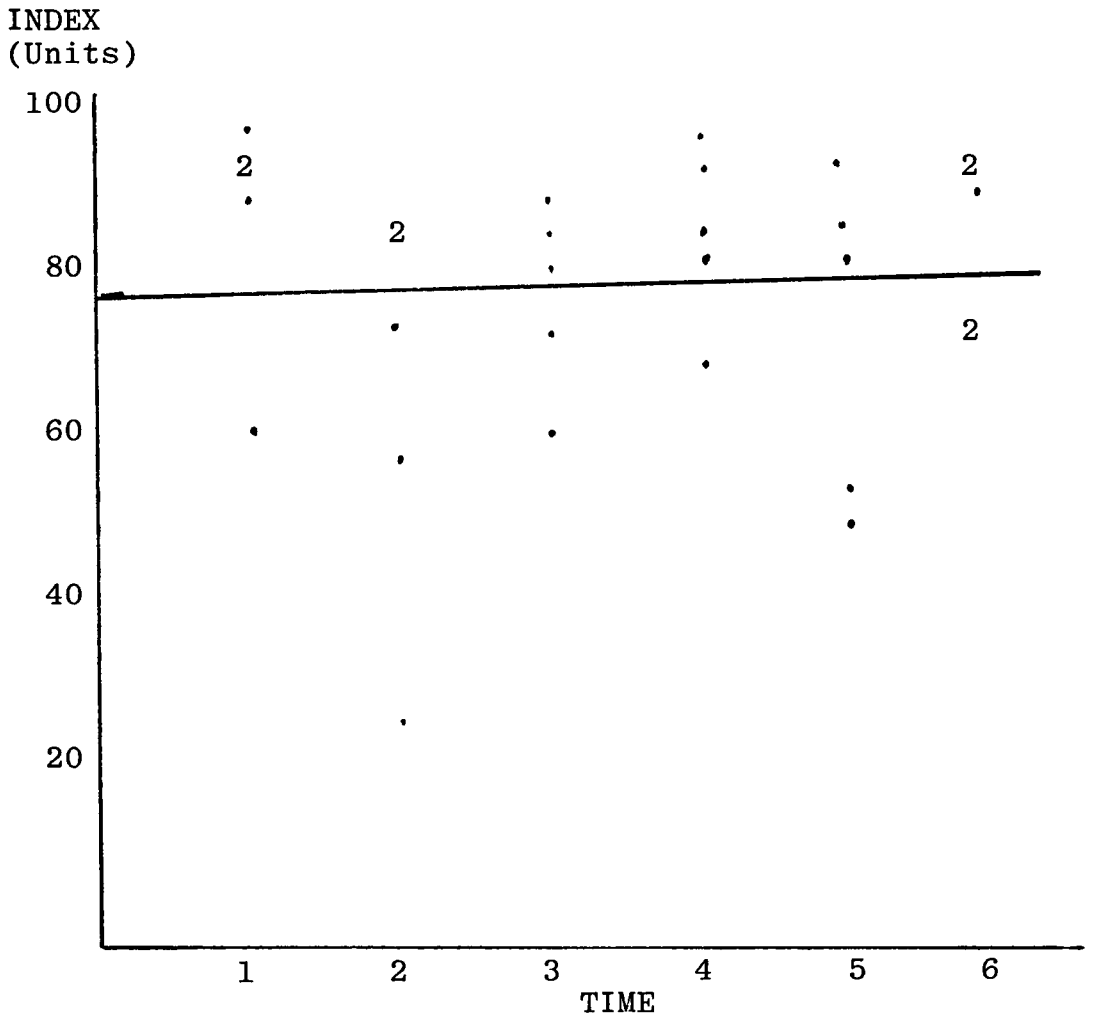


FIGURE 10. Test Reliability for the Control Group.

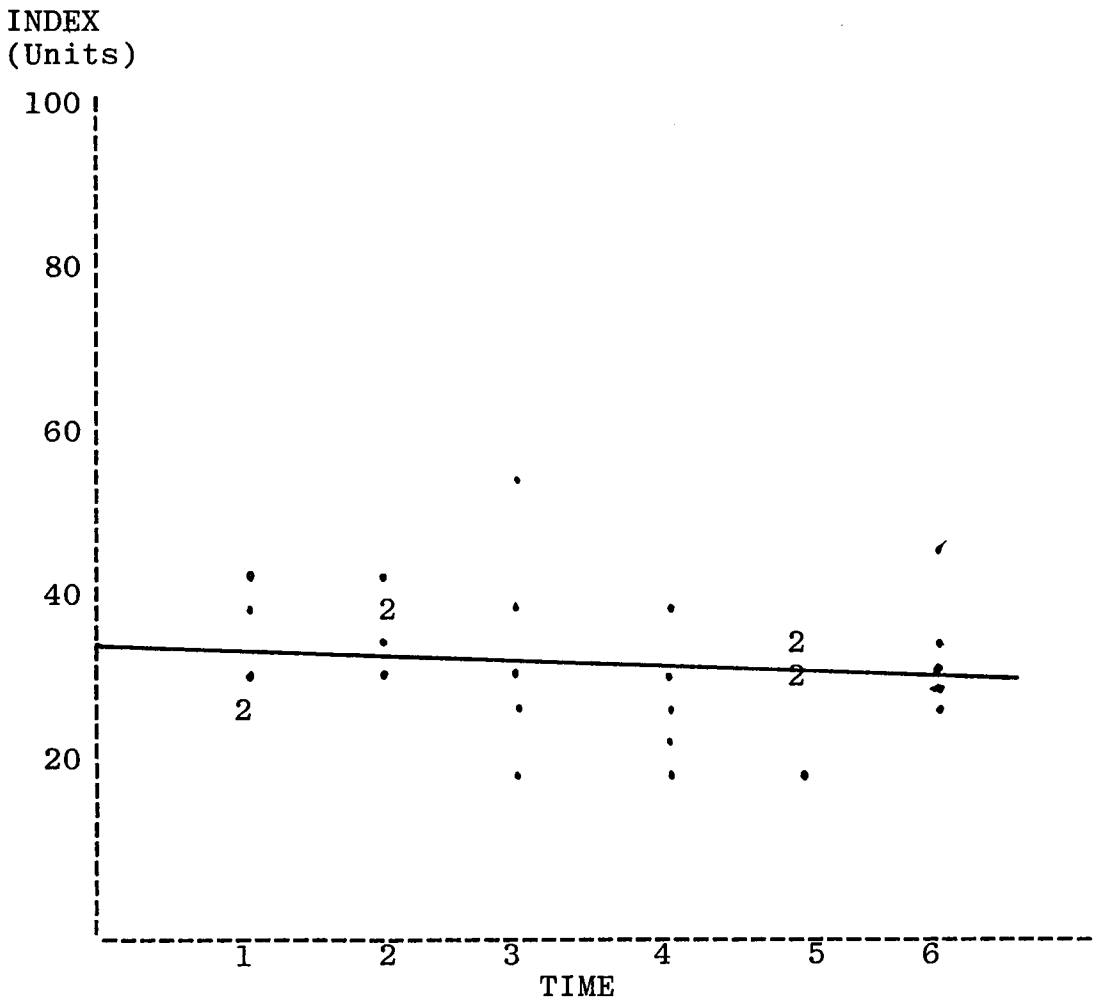


FIGURE 11. Item Difficulty for the Experimental Group.

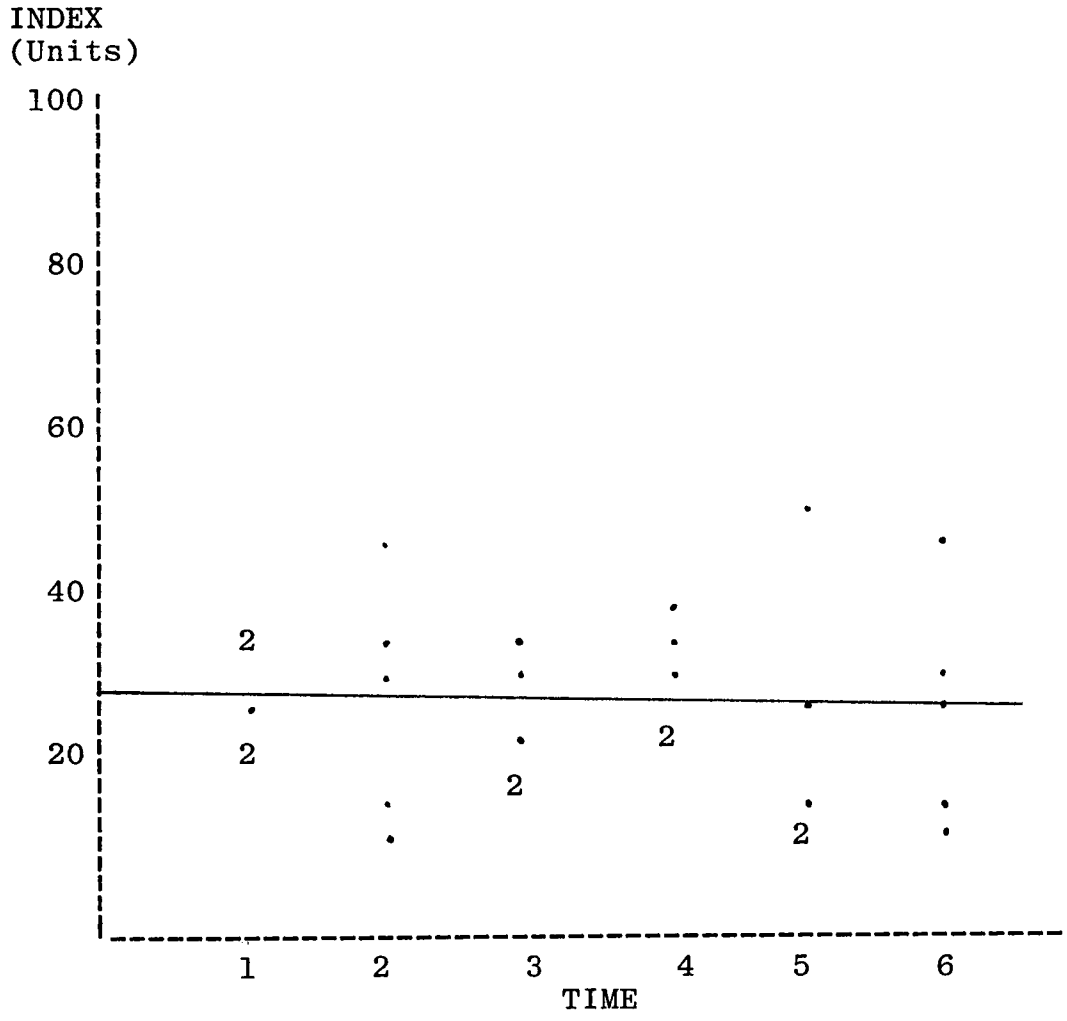


FIGURE 12. Item Difficulty for the Control Group.

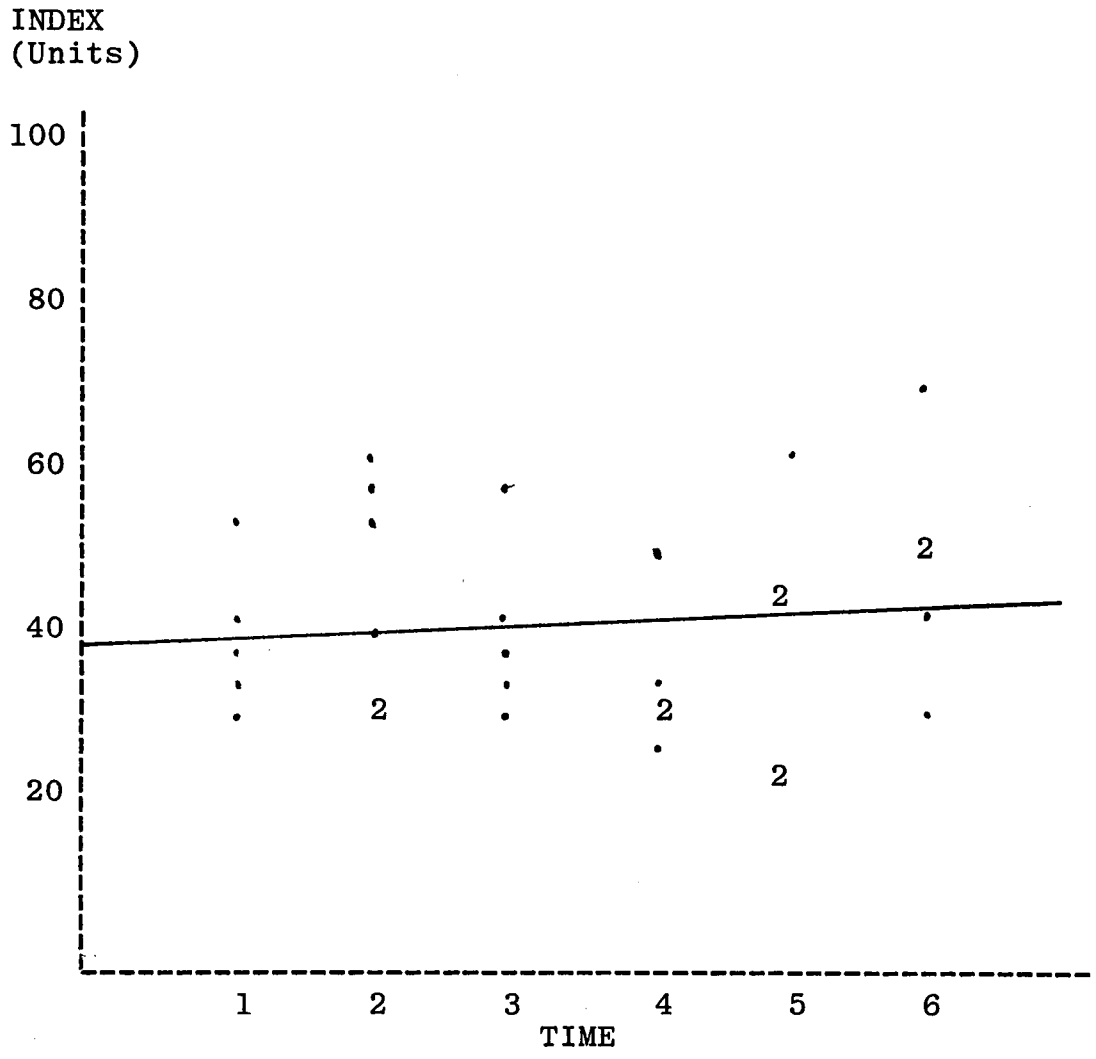


FIGURE 13. Item Discrimination for the Experimental Group.

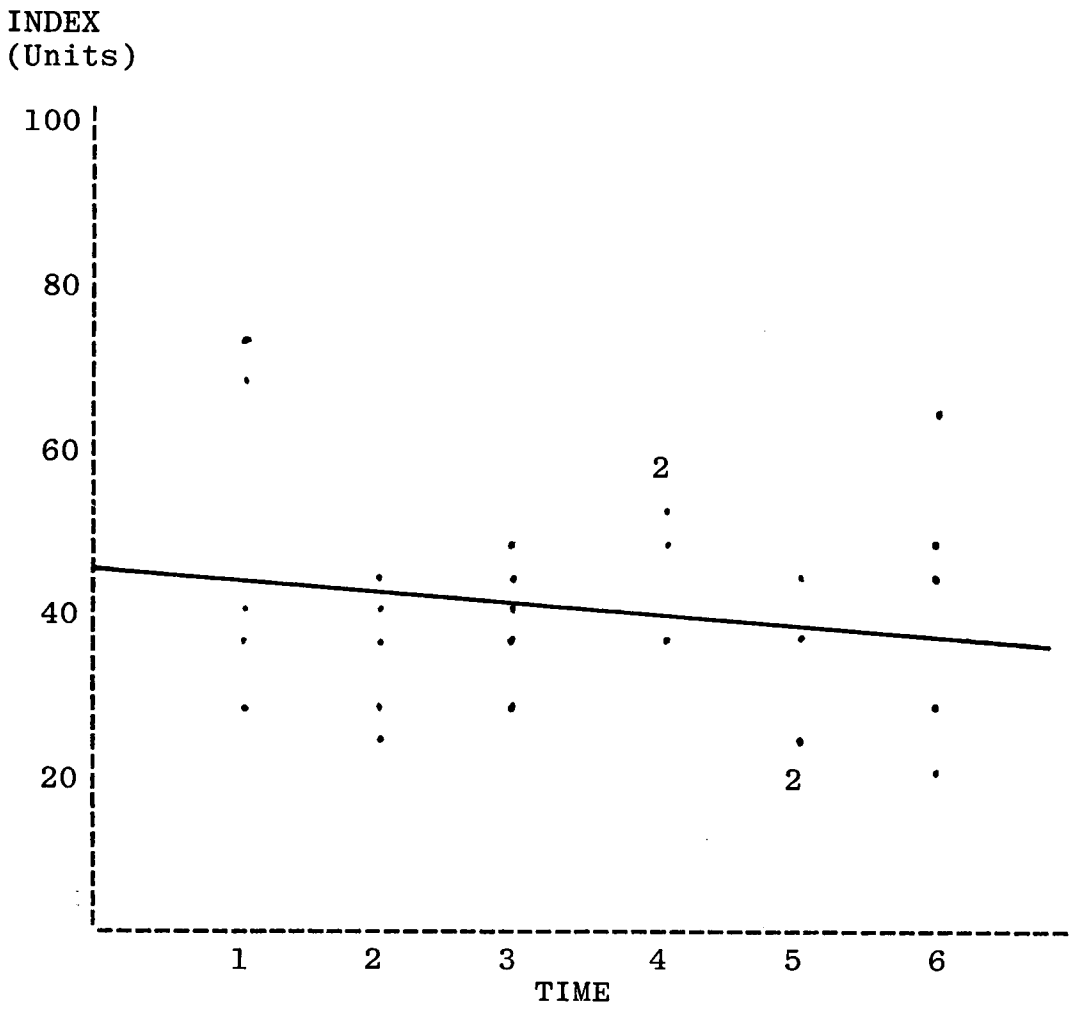


FIGURE 14. Item Discrimination for the Control Group.

APPENDIX F

Turn this side up to show the answer as: TRUE A 1

Turn this side up to show the answer as: FALSE B 2

Turn this side up to show the answer as: C 3

Turn this side up to show the answer as: D 4

VITA

Canton Livingston Napier was born in Myrtle, Kentucky on January 13, 1937. He attended elementary school in Martin, Kentucky and was graduated from Martin High School in 1955. He entered the United States Air Force immediately upon graduation.

Receiving his discharge in 1959, he entered Union College, Barbourville, Kentucky. In August, 1962, he received a Bachelor of Arts degree from Union. He began his teaching career in the Oak Ridge City Schools as a social studies teacher in a junior high school.

He received the Master of Science degree with a major in Educational Administration and Supervision from The University of Tennessee, Knoxville in August, 1971. In August, 1979, he received the Doctor of Education degree with a major in Educational Administration and Supervision from The University of Tennessee, Knoxville.

He is a member of Phi Delta Kappa, Kappa Delta Pi, Iota Sigma Nu, and Pi Gamma Mu. At present, he is a life member of the National Education Association.

He is married to the former Ellen Ruth Buchanan of Benham, Kentucky. They have a daughter, Lisa Lyn.