

1-1-1977

Improving the instructional use of test data through an in-service staff development program for educators.

Charles Joseph Clock
University of Massachusetts Amherst

Follow this and additional works at: https://scholarworks.umass.edu/dissertations_1

Recommended Citation

Clock, Charles Joseph, "Improving the instructional use of test data through an in-service staff development program for educators." (1977). *Doctoral Dissertations 1896 - February 2014*. 3381.
https://scholarworks.umass.edu/dissertations_1/3381

This Open Access Dissertation is brought to you for free and open access by ScholarWorks@UMass Amherst. It has been accepted for inclusion in Doctoral Dissertations 1896 - February 2014 by an authorized administrator of ScholarWorks@UMass Amherst. For more information, please contact scholarworks@library.umass.edu.

UMASS/AMHERST



312066013294687

IMPROVING THE INSTRUCTIONAL USE
OF TEST DATA THROUGH AN
IN-SERVICE STAFF DEVELOPMENT PROGRAM
FOR EDUCATORS

A Dissertation Presented

By

Charles Joseph Clock, Jr.

Submitted to the Graduate School of the
University of Massachusetts in partial fulfillment
of the requirement for the degree of

DOCTOR OF EDUCATION

November 1977

Education

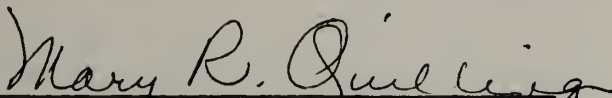
IMPROVING THE INSTRUCTIONAL USE
OF TEST DATA THROUGH AN
IN-SERVICE STAFF DEVELOPMENT PROGRAM
FOR EDUCATORS

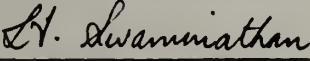
A Dissertation Presented

By

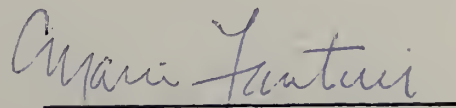
Charles Joseph Clock, Jr.

Approved as to style and content by:


Mary R. Quilling, Chairperson


H. Swaminathan, Member


Richard F. Haase, Member


Mario Fantini, Dean
School of Education

ACKNOWLEDGEMENTS

A project of this type requires considerable support from those who are involved in its implementation. I would like to express my sincere appreciation to all the educators who took the time to participate in this study and who contributed so much to its fulfillment. A very special acknowledgement also to my committee. Mary Quilling and H. Swaminathan have helped me appreciate the value of higher education and its relevance to my professional improvement. They are excellent teachers whose value to their profession includes and extends beyond the classroom. Dick Haase has provided excellent support and constructive criticism -- introducing an element of objectivity which has strengthened the entire project. My parents have also contributed much needed support over the years in all my academic pursuits. It would be difficult to count the ways. My wife, Dorothy, deserves special credit. Without her support and sacrifice, this project would never have been started and certainly would never have been completed.

ABSTRACT

IMPROVING THE INSTRUCTIONAL USE OF TEST DATA
THROUGH AN IN-SERVICE STAFF DEVELOPMENT
PROGRAM FOR EDUCATORS

(February, 1978)

Charles J. Clock, Jr.
B.A. & M.S. San Jose State University
Ed.D. University of Massachusetts

Directed by: Dr. Mary R. Quilling

This study was concerned with the development and field testing of an in-service staff development program designed to upgrade educators' skills in using standardized test data. The field testing of the staff development program was accomplished in five school districts involving a mix of urban and suburban populations and administrator, teacher, and educational specialists as program participants.

An adaptation of the Provus Discrepancy Evaluation Model was used in developing and evaluating the staff development program. The emphasis was placed on a formative process of program development and evaluation. Pre- and post-assessment instruments were used before and after training to assess possible changes in cognitive skills of participants and their attitudes toward tests and measurement. Questionnaires and follow-up interviews were

also used after training to examine program effectiveness and aid in program review and modification.

The results indicated that the formative processes of program development and evaluation used in this study were effective in creating an in-service staff development program in tests and measurement. The majority of the participants reflected positive attitudes toward the program. Pre- and post-assessment of cognitive skill information showed significant increases in tests and measurement skills. Future studies on the use of this staff development program are recommended to determine long range instructional effects and develop further follow-up training activities relevant to local educational needs.

T A B L E O F C O N T E N T S

	Page
ACKNOWLEDGEMENTS	iii
ABSTRACT	iv
LIST OF TABLES	viii
 CHAPTER	
I INTRODUCTION	1
Problems in Test Use	2
Purpose of the Study	6
Definition of Terms	7
Methodology	9
Significance of the Study	11
Organization of the Dissertation	11
II REVIEW OF THE LITERATURE	13
Importance of Testing	13
Attitudes of Educators Toward Testing	19
Need for Teacher Training in Tests and Measurement	22
Current Approaches to Staff Development in Tests and Measurement	28
Need and Rationale for Program Development	34
General Program Considerations	34
Specific Content Considerations	39
III DEVELOPMENT, IMPLEMENTATION AND EVALUATION OF A STAFF DEVELOPMENT PROGRAM ON TESTS AND MEASUREMENT	43
Design Stage	47
Statement of Problem	47
Problem Verification and Needs Assessment	47
Program Specifications	48
Program Design	55
Evaluation of the Design Stage	57
Installation and Process Stage	58
Program Implementation	59
Evaluation of the Installation and Process Stage	62
Instrument Design and Development	63
Product Stage	67

CHAPTER	Page
IV RESULTS	69
Evaluation Processes and Results	70
Design Stage	70
Installation and Process Stage	75
Product Stage	88
V CONCLUSIONS AND RECOMMENDATIONS	95
Summary of Findings	96
Delimitations of the Study	101
Training Considerations and Recommendations	104
Further Development and Research	107
BIBLIOGRAPHY	110
APPENDICES	115
A Pre- and Post-Assessment Instrument	115
B Questionnaire Form	124
C Interview Form	130
D Pre- and Post-Assessment Results of the Non-Cognitive Skill Test Items	133
E Quantitative Results of the Interview	137
F Quantitative Results of the Questionnaire	139
G Changes in Program Content as a Result of Participant Feedback	142
H Final Version of the Staff Development Program	145

L I S T O F T A B L E S

		Page
TABLE		
1	School Districts Participating in the Staff Development Program on Tests and Measurement.	60
2	Content Categories and Item Numbers Included on the Pre/Post Assessment Instrument	66
3	Percent of Correct Responses in Each of the Cognitive Skill Areas as Measured by the Pre- and Post-Assessment Instrument	76
4	Pre- and Post-Assessment Results on Cognitive Skill Information About Tests and Measurement.	78

C H A P T E R I

INTRODUCTION

Most people who have attended school in this country, and many who have applied for employment have experienced the social consequences of testing. Testing has directly affected people's lives in their search for knowledge, academic opportunities, and vocational choice. Goslin (1967) estimates "that each of the fifty million or more school children in the United States takes, on the average, three standardized tests per year." An activity of such magnitude and importance requires considerable care in its development and application.

Although reports of using measurement of performance for selection purposes date back as early as 2200 B.C. in China (DuBois, 1964), testing, as it exists today, is a relatively new activity. Systematic attempts to observe and record data for experimental purposes did not begin until the early 19th century, and these attempts were in the biological sciences (Tuckman, 1975). By 1905, Alfred Binet and Theodore Simon produced the first measure of "human intelligence," which was revised in 1916 at Stanford University and became known as the Stanford-Binet

Intelligence Scale. Cronbach (1970) and Tuckman (1975), in their separate reviews of test development history, show that large scale test development activities of human behavior and performance began around 1920. There was considerable activity in the development of achievement tests during the 1930's and 1940's. Since the 1940's, test development activity has focused on creating and refining a wide variety of instruments and techniques to measure many facets of human behavior.

The growth of test development activity and the proliferation of testing have more recently been stimulated by the growth of the computer industry. The computer has not only contributed to the ease of scoring and reporting test data, but has aided test publishers by providing greater speed and more sophisticated analytical techniques for test development.

While the testing industry has made great strides in a relatively short period of time, there is still much to learn about the intricacies of measuring human behavior. However, the point in time has been reached where relatively sophisticated instruments are available for educational measurement. More important now is the issue of test use.

Problems in Test Use

One problem related to the use of test data in the

educational environment has to do with the perceptions of educators as to the purposes of testing. Tests, particularly the norm-referenced type, are designed to be used for selecting children for various educational activities, placing children into ability groups or an instructional sequence, diagnosing general skill deficiencies, and measuring individual or group academic progress as compared with that of a defined reference group. While these purposes of testing are broad, the perceptions of educators about the purposes of testing are frequently more limited. In the educational environment, many educators, particularly administrators, see testing solely as providing the final measure of their educational products. Norm-referenced standardized test data are often perceived as serving the singular function of providing tangible evidence of the success or failure of an instructional program, school or school system, out of the context of any other information. This summative emphasis on the use of test data makes testing threatening to professional educators and tends to diminish other valid and more useful purposes.

The multiple purposes of testing and their contributions toward providing a variety of useful instructional information to educational decision-makers are emphasized by Dyer (1971) when he stated:

I am convinced that, by and large, standardized achievement tests, despite their admitted defects

have, over the years, contributed in a major way to the improvement of American education and that they have great potential for bringing about further improvement -- provided that their limitations are understood and that they are seen by everyone concerned in the perspective of the feedback model of evaluation. (p. 14)

Tests can yield information that will enable educational decisions to be made. However, in addition to understanding the purposes of testing, educators also need to understand the limitations of tests and the results they provide. For example, tests are restricted to the specific domain they were designed to assess. A typical norm-referenced standardized achievement test is designed to measure the basic skill content areas of reading, language arts and mathematics, and to assess only a sample of specific skills in these areas. They are also given at specific times of the year and these times may or may not correspond directly to the scope and sequence of the content of the operational instructional programs. Norm-referenced standardized achievement tests also have psychometric qualities which will cause them to produce predictable results with certain segments of the population or children with particular socio-economic background characteristics. Professional educators who accept these limitations of testing are in a position to place testing in a better perspective.

Another problem in the use of test data relates to educators' ability to use test information. A study

conducted by Brim, Goslin, Glass and Goldberg (1964) was one of the first to reveal the magnitude of the national testing movement and the associated lack of testing knowledge and training on the part of the test users. Their study revealed that despite the large number of commercial achievement tests administered each year (approximately 135 million), the majority of teachers have had no training in tests and measurement since they left college, where perhaps they had one course in this area. A later study by Mayo (1970) attested to the lack of formal teacher preparation in tests and measurement in college and a continual lack of information two years beyond graduation. What little training teachers were found to have was usually in the interpretation of individual pupil test data with the emphasis on learning the definition of terms rather than the development of any technical background in tests and measurement.

The results of the study by Brim et al. (1964) revealed that test use was more a function of teacher knowledge about tests than any of the other training variables they examined. Teachers who had been given a firm technical background in the basic principles of measurement made better use of test results and at the secondary level, demanded greater help from counselors in the area of test interpretation and use of the results. These findings

corroborated those of Hastings, Runkel and Damrin (1961), who found that teacher attitudes and practices toward testing could be positively influenced with an intensive summer training program on the use of test results.

Purpose of the Study

Frequent assumptions made by test scoring agencies, educational researchers and evaluators, and local school personnel in releasing test data are that (a) the results will be useful to teachers and administrators in improving instructional practices, (b) the presentation of results will reflect and be sensitive to the aspects of the data that could significantly effect instructional decision-making, and (c) the presentation of the results is in a form and context that educators and the public can easily understand and interpret correctly. These assumptions, though important, are not always realized in the educational environment. Undergirding this study is the intent to provide an instructional program so that such assumptions can be realized.

The purpose of this study will be to develop and field test a staff development program that is effective in upgrading educators' skills in both understanding and making decisions based on standardized test data. It is presumed that such a program will increase the instructional use of

test information by both school administrators and classroom teachers.

Definition of Terms

Evaluation, testing, measurement, and assessment are terms that are frequently used synonymously. For the purpose of this study, it is necessary that these terms and others be expressed in a simplified manner, placed in the proper perspective, and their relationship to one another defined as follows:

Measurement: The act of observing a behavior or characteristic of an individual or group and recording the information usually, but not always, in numerical terms.

Testing: A type of measurement activity in which specific instruments are used to determine individual or group performance characteristics.

Assessment: A comprehensive process involving the specification of a problem, the identification of variables that can affect the problem, and the use of measurement techniques to gather information for the evaluation process.

Evaluation: A judgmental process applied to the results of assessment for instructional decision-making.

Formative Evaluation: The process of gathering information during the course of an instructional activity for the purpose of reporting this information to instructional planners, developers and implementors to revise and improve the activity while it is in operation.

Summative Evaluation: The process of gathering information during or after the course of an instructional activity to determine the activity's

overall effectiveness. The results are generally terminal -- in the form of a final report at the conclusion of an instructional activity.

Norm-Referenced Test: An instrument that primarily compares the performance of an individual on a given task with that of a defined group on the same task. The results are scores which are used to measure the relative standing of an individual in a group.

Criterion-Referenced Test: An instrument that primarily describes the relationship between an individual's performance on a task and the nature of the task itself. The results are an absolute measure in that they can reveal discrepancies between actual and desired performance in specific skill areas.

An important distinction in the above definitions is the difference between the terms "testing" and "evaluation." These two terms are the most frequently misinterpreted and misused. It is possible to give a test and not evaluate, and it is also possible to evaluate without giving a test.

Dyer (1969) has further defined evaluation as "A process for reaching decisions about the total educational program and its numerous components on the basis of relevant, dependable and interpretable information about students, the condition of their learning, and the actual events that take place in the classroom." The keys to this definition are the quality of the information and the wisdom with which the information is applied. Judgments made about an individual, a group or an instructional process should be based on appropriate measures for which there is knowledge and understanding about their applicability to a particular question. How

effectively these kinds of data are applied to a particular problem depends largely on how the results are analyzed and presented to the user. If testing is a part of an evaluative process, then it is important for the users of test data to be aware of the implications that these data may have in the more complex structure of evaluation.

Methodology

The study was descriptive in nature and focused on the development and field testing of a staff development program for educators in the area of tests and measurement. A systems approach was used in the design, development and implementation of the program. An adaptation of the Provus Discrepancy Evaluation Model (Provus, 1971) was used as the basis for a formative evaluation process. This model was chosen as a basis for evaluation due to its apparent relevance as both a developmental and an evaluative model.

The emphasis on the formative aspects of program evaluation was considered vital to the program's overall effectiveness. Summative information was obtained through follow-up questionnaires on both cognitive skills and attitudes, and interviews conducted with educators who participated in the program.

The design and implementation of the staff development program commenced in the fall of 1976. Approximately 150

teachers and administrators in five New England school systems agreed to participate in the staff development program. Their agreement to participate in the program involved a commitment to assist in the formative evaluation process. They agreed to participate in a survey administered anonymously both prior to and after training that elicited their knowledge and attitudes about tests and measurement. A random sample of 80 teachers and administrators were interviewed following the program to determine their understanding and use of test data in instructional decision-making and the types of information they felt were pertinent in their work with children. All participants in the staff development program were asked to complete a questionnaire during the month of February describing their use of test data and their attitudes toward testing. In addition, all participants were encouraged to make critical comments, both oral and written, during and after the implementation of the program. These comments were used as one basis for revising the program prior to its subsequent administration in a different school district. The program was implemented between October and December of 1976, and was designed to be completed in a time span of from six to eight hours in order to meet varying local time constraints and desired levels of involvement.

Significance of the Study

The value of this study lies in the contribution the staff development program can make to the larger educational community. A primary consideration in the accomplishment of this study is the construction of a product which can be used by other school districts. A staff development program is the product, and it is designed to be flexible enough to be easily adapted to any local testing environment.

Organization of the Dissertation

Chapter II presents a review of the literature reflecting (a) the need for a staff development program on tests and measurement, (b) rationale for course content, (c) the state of the art regarding teacher attitudes and competencies and existing training programs in tests and measurement, and (d) additional ways of analyzing and presenting test data to make them more "productively descriptive."

Chapter III describes the steps involved in program development based on the review of the literature and the feedback process employed during implementation. This will include a detailed description of the system's design, the procedural steps in development and the mechanism for program review and modification. Several questions of both a

formative and summative nature are listed and the techniques used in answering them are identified.

The results of the program development process, including specific modifications, are discussed in Chapter IV. Chapter IV also contains the results of the testing, questionnaire and interview procedures and will include the implications of the informal feedback process.

Chapter V presents the conclusions of the study and any implications for further revision and/or implementation.

CHAPTER I I

REVIEW OF THE LITERATURE

Importance of Testing

The age of accountability has increased the importance of testing in the educational environment. The advent of increased computer technology has made testing more amenable to the demands of both the public and the educators. The demand for information and the availability of it has made achievement testing a prime source of information regarding the success or failure of instructional programs or approaches. While the tests and their related information have become more sophisticated, this rapid growth in technology has not been paralleled by an equal growth in teacher training or expertise in the skills necessary to make effective use of the data. This is particularly true of group performance data from norm-referenced standardized achievement tests which are presented to educators and the public with the assumption that the clients have the basic statistical expertise in order to make valid decisions based on the information provided.

The application of test results has not always yielded positive responses from either the educational establishment

or the public. In fact, the National Education Association has formally requested a moratorium on all standardized tests until the completion of a critical appraisal, review, and revision of current testing programs (Resolutions and other actions, 1972). Standardized testing has received a more recent condemnation by the National Education Association in an article by McKenna (1977), an officer of NEA. McKenna reports that regardless of what is done to improve the interpretation and use of test data, standardized tests are still inherently inadequate and should not be used. Bhaerman (1977), on the other hand, reflects a different perspective from the American Federation of Teachers:

Standardized tests must be kept in proper perspective. Rather than lash out indiscriminately against all such tests, the more sensible route should be to identify specific weaknesses and improve them In identifying the enemy let us keep in mind that in this case the major one probably is misuse and abuse. Tests should not become the main reason for existence but neither should they be totally discarded. (p. 14)

Despite these reports, schools and school districts are being "evaluated" every year by both educators and the public on the basis of data from standardized achievement and aptitude tests. Regardless of the test and measurement skills possessed by educators, test results are playing a major role in educational decision-making. Not only are educators using test results for making major instructional decisions, but parents, boards of education and the non-parent voters are also making decisions about local instructional

practices based on test results. In a study sponsored by the National School Board Association (NSBA), over 50% of the school board members reported that they used test results for judging the effectiveness of district-wide programs and in making decisions about curriculum changes (National School Board Association, 1977). However, only 49% of the school board members claimed they understood test results, and only 53% felt their district school administrators were capable of interpreting the meaning of test scores.

The press is anxious to publish articles revealing test scores, particularly if they contribute to a political concern. These published scores are usually in the form of average scores which frequently lead to further misuse of data. Average scores are limited in the information they reveal and the public generally does not have the educational background or the expertise to interpret them properly. In addition, a study by the National School Board Association (National School Board Association, 1977, p. 26) showed that 82% of the respondents felt news reporters do not understand test results and, consequently, do not interpret them properly to the public. The New York Times (Sunday, May 1, 1977) had an entire section on testing -- most articles pointing out the problems and pitfalls of using norm-referenced standardized tests as measures of pupil performance. The size and extent of the New York

Times article gives support to the importance currently being placed on testing in the public schools.

Further indication of the significance testing has on public education is the national concern over the declining College Entrance Examination Board's Scholastic Aptitude Test (SAT) scores. The national decline of these scores is seen in many educational communities as direct evidence of the failure of modern educational practices. Harnischfager and Wiley (1975) summarized the score decline problem as it relates to the most popular standardized tests. Their study emphasized three major points: (a) there are general score declines and the reasons are not clearly definable, (b) there is a lack of comparative analysis of the tasks, content and scaling involved in different tests as well as a lack of demographic and item/skill data which would make the identification of specific problem areas possible, and (c) there is a pressing need for an overall valuation of tests and testing to determine if there is a match or diversity of test content and school curricular and/or if this relationship changes over time. These points relate to the restrictive nature of the average score as a diagnostic index of performance, the need for examining other types of data as useful indicators of performance, and the basic issue of the appropriateness of the instruments and the way they are used. This latter point was also mentioned by Cronbach (1970) who discussed

the value of considering the application of the results as a prerequisite to decisions regarding the most effective type of item format. Bloom (1970), Popham (1971), Hambleton and Gorth (1971), Glaser (1963), Randall (1972), Ozenne (1971), Sax (1974), and Smith (1973) are others who raised questions about the legitimacy of the types of tests currently used to accomplish certain objectives.

Airasian and Madaus (1976) studied the question of the sensitivity of measures of school and instructional program effectiveness. Their primary concern was the sensitivity of norm-referenced standardized tests in measuring the impact of an instructional approach on groups of children. Four general findings emerged from their study:

1. The use of the total test score in school comparisons hides unique and statistically significant school achievement differences at the item or objective level.
2. The nature of the subject matter tested, independent of the particular type of test used to measure achievement, appeared to effect the magnitude of observed school achievement differences.
3. The psychometric nature of the items comprising a test did not appear to be a central factor influencing the discovery of school achievement differences.
4. The unit of analysis appeared to influence the amount of achievement variation observed. Analysis performed at the school or program level across individual teachers is not sensitive to the achievement variation discovered at the individual teacher level. (pp. 259-260)

The Airasian and Madaus (1976) study raises some interesting

concerns related to the issue of using group test data as a criterion of program effectiveness. Their conclusions indicate that the question is not so much what kind of test is preferable, but how it is used and analyzed. More specifically, a major concern is in how effectively educators and/or evaluators identify the problem to be investigated and apply the proper methods and techniques in determining program effectiveness.

The primary limitation [in studies of differential school or program effectiveness] is the failure to conceptualize adequately the nature of the differences sought, the level at which they are likely to be manifested, and the processes underlying them. The important issue of sensitivity resides not in the tests used to measure achievement, but in the manner in which the problem is conceptualized. (p. 278)

Cooley (1971), Co-director of the Learning Research and Development Center, gives credence to the compatible application of locally developed and norm-referenced tests when he states:

Center-developed [local] tests are important because they answer the question of whether its instructional program actually teaches the behavior it is designed to teach. But a comprehensive evaluation effort needs to do more than that. It needs to demonstrate how well children from the program are equipped to cope after they leave the school. If primary factors of abilities and motives are good predictors of success and satisfaction as young adults . . . and if these factors can be estimated by a mixture of standardized tests and measures derived from operating the instructional model, then these factors can and should be criteria of the program's effectiveness. (p. 22)

The intent of this study is not to dwell on one

measurement methodology versus another, but to define and help resolve the problems associated with the use of tests by educators and specifically the interpretation and application of test results. All of this points to a need for a better understanding of the strengths and limitations of testing, and a general lack of knowledge about the interpretation and application of test results.

Testing is important only if properly applied, and to be properly applied it must be understood. Ebel (1967) points out that the test score reports only the level of knowledge the pupil possesses, not how frequently or how effectively he makes use of it. There is a whole battery of other factors that influence a child's or a group's academic performance which are not a part of the test score. If used in conjunction with other relevant information about pupils, test results can help teachers teach and learners learn by determining the knowledge pupils have to perform designated tasks.

Attitudes of Educators Toward Testing

The data on attitudes toward testing in the educational community are not consistent. Brim, et al. (1964) conducted a survey of 1,754 teachers regarding their attitudes and use of standardized achievement tests and discovered that between 30% and 40% of the teachers felt a fairly great amount of

weight should be given to standardized measures of intelligence and achievement. Goslin (1967) indicated a higher regard on the part of educators for standardized tests than is reflected by such publications as the NEA resolutions on testing (Resolutions and other actions, 1972) which called for a national moratorium on testing. Relatively positive teacher attitudes toward standardized tests were also reported by Short and Szabo (1974) who found an associated low level of knowledge about tests. They discussed the possibility of improving the knowledge base and capitalizing on positive attitudes but did not report any studies on the relationship of gains in knowledge to gains in attitude. Cormany (1974) reported that those educators who considered themselves well informed about tests and the school testing program had significantly higher attitude scores than those who did not consider themselves as well informed. Hastings, et al. (1961) found that teachers who had a firm technical background in the basic principles of measurement were more accepting of tests than teachers who lacked this expertise. Hotvedt (1974) discovered that in addition to having an adequate knowledge base about testing, an equally important factor affecting the attitude of teachers toward testing is the availability of results and the time teachers have to use them. If the results are available early in the school year and teachers have the proper training to use them, the probability is greater that they will be used for more

effective planning, placement and instructional purposes.

Teachers' attitudes about testing also appear to be influenced by the support administrators give to the activity and the quality of the testing program. Stuck and Wyne (1977) examined a district-wide testing program in North Carolina where they discovered little administrative support, poor match of tests with curricula, no in-service training on tests and measurement, and ignorance on the part of the pupils as to the use or feedback of the results. Not too surprising, the attitudes of both teachers and pupils toward testing was found to be relatively negative. Teachers urged the establishment of a Testing and Research Coordinator, examination and modification of the entire testing program, and in-service training on tests and measurement for those who have to interpret test results. It is noteworthy that despite all the problems associated with their testing program, the cry was for improvement and training rather than abolishment.

The attitudes of educators about tests and testing and their willingness to use them appears to be related to how much they know about them. This need for an effective knowledge base is not limited to the classroom teacher. Traxler (1967) described the use of test results as "an all-faculty function. When it is accepted as such, pupils and teachers alike can benefit greatly from a comprehensive, regular, systematic testing program."

Need for Teacher Training in Tests and Measurement

In 1964, the Russell Sage Foundation investigated the use of testing in the United States. Their results showed that at that time, 45 million school children were taking an average of three commercially produced tests each year, or approximately 135 million tests taken per year (Brim, et al., 1964). Of the teachers contacted in the Northeast, 27% felt the single most accurate measure of a student's intellectual ability was provided by standardized achievement test scores. Between 40 and 45% of these teachers felt standardized achievement tests should contribute a fairly great or great amount toward recommending students to colleges, taking extra courses, occupational counseling and college selection. Approximately 60% felt they should be a primary source for assigning students to special classes. However, in contrast with the teachers' stated desires for testing, 50% had never administered a standardized achievement test and 67% reported never having attended a clinic or workshop on the content, philosophy or methodology of standardized testing outside of their college experiences. Only 50% had had more than one course in college on tests and measurement and 22% had none. The statistics for teachers taking courses in methods of research revealed an even greater percentage of non-exposure. Goslin (1967) found similar results where teachers in the elementary

schools were heavily involved in the administration and interpretation of test data, but almost half of those interviewed had never had any formal training in tests and measurement.

It may be argued with considerable logic that it is not necessary for an elementary or secondary school teacher to have had a formal course in tests and measurement in order to be able to administer a standardized achievement or intelligence test to a group of pupils The fact that he possesses his pupils' scores on standardized tests, however, places the teacher in the position of being more than a mere test administrator. (p. 127)

Hastings, et al. (1961) explored the question of teacher competency in tests and measurement and reached two basic conclusions. They learned that test use was more a function of technical knowledge about tests than any of the other kinds of training studied. Secondly, they discovered that as teachers learned more about tests, they tended to make more and better use of them, and secondary teachers demanded greater help from counselors in the area of test use and interpretation.

Aside from the findings of the present research, our experience with such a program in the Fairview School of Project 509 suggested that learning technical concepts and principles was a necessary pre-condition for later and wider application of the skills acquired. (p. 211)

Mayo (1970) demonstrated that if teachers are to make better use of test results, then they must be given more pre-service and in-service training in the interpretation and use of standardized test results. Fleming (1971) also

observed the lack of teacher training and stated:

Evaluation techniques probably have received less emphasis than any other facet of the teaching-learning environment. Few teachers have arrived in the classroom with developed skills in designing observation guidelines, constructing classroom tests, and implementing objective marking models let alone interpreting results of standardized tests. (p. 71)

Fleming continues by indicating several trends in the direction of an improved evaluative process in schools utilizing standardized tests. These trends include:

1. Recognition that in-service and pre-service activities for teachers in the evaluative process pertinent to instructional assessment are priorities -- whether standardized tests are used or not.
2. Recognition that effective testing programs require upgrading competencies of the school staff in administration of tests, interpretation of results, and dissemination of information. (p. 72)

All these studies point to the lack of training in measurement and testing on the part of educators. This lack of a sufficient teacher knowledge base about testing and the contrasting abundance and variety of testing information about individual pupils and instructional groups could foster either the misuse or lack of use of test results. More recent studies, such as the one by Hotvedt (1974) also show that teachers' knowledge of tests and measurement is relatively low when considering the amount of emphasis placed on these kinds of results.

Brady (1977) reports that:

Probably few teachers emerge from teacher preparation programs well equipped as evaluators or consumers of evaluation information. As professional preparation units are whittled away, this is an area of content often omitted. (p. 5)

Teachers, themselves, are urging the development of pre-service and in-service training in tests and measurement, particularly test interpretation (Bhaerman, 1977). The American Federation of Teachers Task Force on Educational Issues states that, "The consensus of opinion [about testing] derived from the AFT survey is that the wisest approach seems to be not to burn down the barn, but to improve the structure" (Bhaerman, 1977). This same Task Force also calls for in-service education for all involved in test utilization.

This desire is not reflected in all teacher organizations. As reported earlier, the National Education Association (Resolutions and other actions, 1972) has strongly encouraged the elimination of group standardized intelligence, aptitude and achievement tests. This NEA resolution was repeated in 1976. A recent NEA journal article (McKenna, 1977) claims that no amount of training can erase the inherent problems associated with standardized tests. It is assumed that the author is referring to norm-referenced standardized tests, since criterion-referenced tests are mentioned as acceptable alternatives.

The problems associated with test misuse or misinterpretation of data are not eliminated with the recommended

NEA sweep of norm-referenced standardized tests. Teachers and administrators still need to understand the basic principles of tests and measurement in order to make effective use of any kind of test information. Many instructional decisions are made on the basis of teacher-made tests -- probably more than are made on the basis of results from large scale district-wide norm-referenced testing programs. This is particularly true where individual children are concerned and where children are functioning in a graded environment. Even the National Education Association (Teacher-made Tests, 1977) encourages the development and use of proper test construction procedures in the creation of teacher-made tests. A recent guide published by the National Council on Measurement in Education (NCME, 1976) has several articles emphasizing the need for teacher training in test construction, interpretation and use as a means of counteracting the negative criticism.

Much of the negative criticism on the part of educators toward testing is probably due to their lack of knowledge about tests and how to use test results. Jackson (1968) quotes a second grade teacher's very limited use of test information when she stated:

The Iowa Test is given in third grade, but the results don't mean anything until the child has taken it again in the fourth grade. You have to wait a whole year before you can tell anything about it.
(p. 124)

This statement, which is certainly not atypical, reflects a

lack of knowledge of how to use test data. It could also reflect a lack of knowledge or support on the part of the local school administrators about how to provide the most effective test data or encourage its use. It may also reflect a lack of sensitivity on the part of the test publisher to provide the information the teacher needs in a way in which it can be readily understood and used. All of these are common problems in local school districts. Whatever the case, this teacher has formed an opinion about a norm-referenced standardized test which is based on ignorance.

The literature reflects a void outside of the college environment in training teachers and administrators about tests and measurement. It is not surprising that educators are asking for help in this area, particularly since the public has made test results a major ingredient in the issue of accountability. The question is not so much one of whether or not there should be staff development in tests and measurement, but what kind of training is most beneficial. The colleges and universities can and should focus both on the theoretical and practical application of test data. However, it is in the classroom with "real" data where the educational practitioner is faced with an immediate need for guidance in the interpretation and use of specific test data. A significant portion of this guidance should be

concerned with the limitations as well as the strengths of testing. It is important for educational practitioners and the community they serve to understand that test results represent only a sampling of human behavior in a specific content area. Test results have meaning, but must not be interpreted out of the context of other information known about a child, groups of children or educational programs they attempt to measure. For this reason, it is necessary for the training of test users to go beyond the classroom. As Dyer (1973) stated:

The field of education has become strewn with politics, and educational testing has become an instrument, if not a weapon, in the political process. And this means that our worries today about the mishandling of tests and the misuse of test scores must embrace not only school personnel, but also politicians and the diverse and pluralistic constituencies they serve. (p. 86)

Current Approaches to Staff Development in Tests and Measurement

Most in-service staff development programs in testing appear to be developed to acquaint educators with a particular instrument being used in a school system. These programs tend to emphasize test interpretation skills and the improvement of student test taking performance. A good example of this type of program is one proposed in Charlottesville, Virginia (Grant, 1976). The Charlottesville Public Schools has submitted a proposal to develop an in-service training

program for teachers in the analysis of test results and the re-evaluation of classroom objectives. The emphasis in the Charlottesville program is on presenting the relevant fundamentals of test interpretation in layman's terms and is oriented toward teacher interpretation of test results and the application of these results to making decisions about curriculum. The Charlottesville program is based entirely on the SRA achievement test series and is divided into two phases. The first phase involves a series of six workshops held in half-day sessions over a period of four months. The second phase involves a continued series of undefined in-service training activities for a period of three years. The initial product of this effort will be a teacher manual which will emphasize the identification of behavioral objectives and content areas measured by the SRA test along with statistics unique to the SRA data which can be used to help teachers interpret their students' test scores. The objectives of the overall training program will be to improve teachers' attitudes toward achievement tests, their ability to interpret test results, and to improve student scores on the SRA test battery. A major emphasis in this program will be on the workshops involving hands-on work with actual pupil data (grades 2 and 4) covering the following areas: (a) interpreting test results pertinent to the SRA Achievement Test, (b) identification of behavioral objectives and content areas of both

the SRA test and the Charlottesville curriculum, (c) location and/or development of materials to alleviate specific skill weaknesses, (d) relating key instructional materials to performance objectives and assessment of mastery levels, and (e) factors affecting test results. Few local school efforts at upgrading educators' skills in tests and measurement are as formally expressed as the Charlottesville proposal. Most local school efforts to upgrade tests and measurement skills appear to be simply extensions of material covered in the test manuals they are using.

Educational Testing Service (ETS) is currently in the process of developing a "core" training package with several content and methodological options. The focus of the ETS program will be on pre- and in-service training in classroom test construction, test administration procedures, and test interpretation to students, parents and community groups. The ETS program is still in the developmental stage, but appears to emphasize test construction as a major component. It will be geared primarily to classroom teachers and administrators.

College and university programs for teacher training in tests and measurement vary widely in course content and emphasis. An example of a very complete and well documented program on the basic principles of testing has been developed by Hambleton (1974) at the University of Massachusetts. Hambleton's program is a one-semester course

covering the following areas:

1. Introduction to testing
2. Descriptive statistics
3. Scores and norms
4. Reliability and validity
5. Selecting and evaluating standardized tests
6. Factors affecting test scores
7. Objective-based instruction, testing and measurement
8. Test construction techniques
9. Measurement of achievement, aptitude and personality
10. Designing school testing programs

One thing that makes this program unique is the documentation which the students have to use as reference material. Each section of the program is well outlined with ample references for further independent study.

In addition to programs such as these, several publications have recently become available which are designed to upgrade educators' tests and measurement skills. For several years, the Test Department of Harcourt, Brace and Jovanovich, Inc. has been publishing testing bulletins referred to as Test Service Notebooks. These publications are offered as a professional service to the educational community and are designed to inform the test users of various aspects of test construction, interpretation, definition and use.

Several programmed texts on tests and measurement have also been developed. Harrington (1968) has produced a programmed text on a basic course in tests and measurement. This course covers definition of terms, development of norms,

processes of standardization, mean, median, standard deviation, and derived scores. Most of these programmed learning materials emphasize the acquisition of basic statistical skills. Some examples of texts which stress the statistical aspects of tests and measurement are STATLAB (Hodges, Krech and Crutchfield, 1975), a simplified programmed text by Amos, Brown and Mink (1965), a series of teacher oriented texts on statistics by Gellman (1973), Bruning and Kintz (1968), Wick (1973), TenBrink (1974), Huck, Cormier and Bounds (1974), and Popham (1975). One trend that is consistent through all these materials is the need for a clear definition of test and measurement terms. This is a major emphasis in Popham's (1975) book. There is also an emphasis in these materials on statistics that are descriptive of the distributions of group data as well as the statistics that relate specifically to the comparative interpretation of individual student performance characteristics.

Lyman (1971) and Kirby, Culp and Kirby (1973) have published excellent resource books for teachers on how to interpret test scores. Lyman's book is based not only on responses from other professionals in the testing field, but on his experiences in teaching test users and their feedback to the matching of material to needs. Both the Lyman (1971) and the Kirby, et al. (1973) books follow a basic pattern of definition of test and measurement

terminology, description of basic statistics used in testing, and the application of this information to the instructional environment.

Current approaches to developing tests and measurement skills in the educational community appear to focus on the methods, techniques and application of performance data for the improvement of instructional practices. The emphasis appears to be sound but the involvement of the educational practitioner is limited. College courses in tests and measurement are frequently either not required or are taught by those with limited knowledge in this area. Pamphlets distributed by test publishers are worthwhile but not usually read by the classroom teacher. Few teachers take advantage or even know about the programmed texts on tests and measurement that are available. In-service training programs that do exist are generally specific to a particular testing program, and it is sometimes difficult for educators to generalize this information to the greater context of information gathering and utilization. The basic materials exist for training educators to make more effective use of test information. What appears to be lacking is the presentation of these materials in an easy to understand and relevant framework, and an effective delivery mechanism for reaching educators at the operational level in their own environment.

Need and Rationale for Program Development

General program considerations

As stated previously, the literature appears to reflect a need for training. Individual teachers as well as groups of educators (Bhaerman, 1977) are calling for training in the areas of tests and measurement. The need exists and must be addressed by teacher training organizations. However, that will not solve the problem of the need that currently exists in the operational educational environment. This need must be satisfied through in-service staff development programs. Initial considerations in program development include: purpose of the program, target group, content, length of the program, types of materials used, and the modes of presentation.

The purpose of the program focuses on the need to upgrade educator's skills in tests and measurement so that testing can be a more effective vehicle for improving an instructional process. Testing is all too frequently either not understood by or seen as a threat to the classroom teacher. A major emphasis in this program will be to discuss the limitations as well as the strengths of testing in an attempt to place testing in the proper perspective. A logical rationale for accomplishing this purpose would be to follow the format developed by Lyman (1971) and Kirby, et al. (1973) where they stress definition of

terms, description of techniques, and application of data to instructional decision-making. This format also appears to be the one used by institutions of higher education in their teacher training programs.

The target group will be classroom teachers since they are the ones who can experience immediate benefit from effective use of test data. If the purpose is to improve instructional practices, then it is the classroom teacher who must be the focal point in the training program. However, school administrators are also significant potential recipients of such training. It is particularly important for administrators to understand the limitations of test data in the process of evaluation. They need to see the power test data can have in the classroom environment; how the teacher can use it to help diagnose skill deficiencies and plan instructional strategies with individuals and groups of children. Administrators tend to get caught-up with the average score as a major index of success or failure. It is akin to not being able to see the forest because of the trees. Administrators can help the classroom teacher, the public, and the efficacy of the whole evaluation process if they have the proper perspective on testing.

Lindvall and Nitko (1975) and TenBrink (1974) have two recent publications on the strategies and content in teacher training program development. Both of these authors

stress the "needs assessment" approach to identifying the strategies for student assessment and the direction for upgrading specific teacher skills in tests and measurement. Staff development programs geared toward more effective use of test results should involve a systematic analysis of the needs local educators have in evaluating instructional approaches or materials. This "needs assessment" process of data analysis can help identify not only the types of data needed for effective decision making, but establish the most effective ways of presenting the data for meaningful interpretation and application. The publications above, the National Council on Measurement in Education (NCME, 1976), Goslin's (1967) assessment of the needs of teachers are only a few examples in the literature that stress particular training needs on the part of educators. These needs tend to focus on defining testing terms (such as types of scores, norms, etc.), basic statistical techniques, and the need for helping teachers interpret test data and their instructional application. This general training approach was also followed by Lyman (1971) who developed his publication based on the needs expressed by both practicing and potential teachers. Consequently, an initial training program was developed which was organized into three parts. Part I was concerned with a definition and clarification of test and measurement terms;

Part II with the definition and application of basic statistical terms; and Part III with the application of all previous information to the interpretation and use of local data to the specific diagnosis and improvement of local instructional practices. The initial program was created to be administered in six to eight hours, depending on local needs.

A major consideration in the design of this staff development program is the extent of training necessary to accomplish improvement in test use. Hastings, et al. (1961) found that a limited number of intensive long-range training programs for the few were preferable in the long run to short "practical" doses for the many. These findings stemmed from the author's theoretical studies of cognitive structure. They report that since the development of a stable cognitive structure is not a linear phenomenon, initial training will serve to reduce cognitive dimensionality by restricting the number of criteria used in judging the problem at hand. After the addition of sizable amounts of training, cognitive dimensionality attains the complexity which is necessary for cognitive stability. Obviously, this would have implications for teacher training, particularly in the operational educational environment where training time is limited. The staff development program is designed as a basic program which can be modified to fit a variety of time frames. It

is recognized that some school systems will not have the flexibility in scheduling staff in-service training programs as others. Consequently, the program is designed as a "core" program with the application section (Part III) being particularly adaptable for workshop-type activities.

The staff development program materials are developed in overhead transparency form with accompanying hard copy. The hard copies are re-prints of the overhead transparencies. This allows for notes to be taken directly on the materials being discussed. Consequently, the mode of presentation is primarily lecture with overhead transparencies for the definition and statistical parts (Parts I and II) and the application phase (Part III) is more amenable to a discussion or workshop format.

The general void of current programs for in-service staff development makes it difficult to contrast this effort with existing programs. A major difference between this and the Charlottesville program (Grant, 1976) is in the content. The Charlottesville program focuses on the SRA Achievement Test Series with the expressed intent on increasing test scores. The focus in this particular program is in making teachers and administrators more aware of the elements of testing, regardless of the types so that the information from tests can be used to improve instructional practices. This is more in line with the philosophy of the ETS program currently under development.

However, the ETS program appears to be more concerned with on-site test construction than generalized principles of test interpretation.

Specific content considerations

A major consideration in this training program is the potential of computer technology to enhance not only the diagnostic aspects of testing, but the nature and presentation of data. The review of the literature is noticeably lacking in information on how to display and interpret group data from tests, and many educational decisions tend to be made on the basis of group data. The need for more effective presentation of group test data is emphasized by Tukey and Wilk (1966) when they stated:

It is insufficient to have results produced; they must be displayed in a manner to satisfy diverse needs of a broad spectrum of individuals Most of us can only appreciate matters with full insight by looking at graphical representations. For large-scale data analysis, there is really no alternative to plotting techniques, properly exploited. A picture is not merely worth a thousand words, it is much more likely to be scrutinized than words are to be read. Wisely used, graphical representation can be extremely effective in making large amounts of certain kinds of numerical information rapidly available to people. (pp. 698 and 700)

The traditional way of presenting group data from norm-referenced standardized tests is by displaying the average as the sole index of performance. These averages may or may not be valid or sufficient as the primary indices of performance for they do not reveal the degree of

variability within or the shape of the distribution from which they are derived. A basic knowledge of the statistics that are involved in score distributions and the limitations of averages should give the educator a better perspective for the interpretation and use of group test data. Klitgaard (1974) urges users of test data to consider other statistics besides the mean (such as standard deviation and skewness) as indicators of group performance. He is careful to caution about the problems associated with using such indices as skewness, but gives strong support for the consideration of the first three moments of the distribution as identifiers of group data rather than the traditional strict reliance on the average. The limitations of the average as a primary index of group performance are discussed by Bloom (1971) in his criticism of educational reliance on the normal curve:

There is nothing sacred about the normal curve. It is the distribution most appropriate to chance and random activity. Education is a purposeful activity and we seek to have the students learn what we have to teach. If we are effective in our instruction, the distribution of achievement should be very different from the normal curve.
(p. 45)

Unless data are obtained on the spread of scores in the distribution or particularly the shape of the distribution, much information will be lost regarding group performance characteristics.

Any effort to upgrade educators' skills in interpreting

and applying test data should involve more than a simplistic description of scores and should emphasize the use of graphic displays. Ladd (1971) reinforces this plea by reporting that teachers are not getting the greatest possible use from test results by simply examining scores. She advocates the use of item analysis, patterns of test item performance, teacher observation of test behavior, and use of graphic results.

Tukey and Wilk (1966) further amplify the importance of making educational data more "productively descriptive." They state the process of data analysis in education is:

To seek through a body of data for interesting relationships and information and to exhibit the results in such a way as to make them recognizable to the data analyzer and recordable for posterity. Its creative task is to be productively descriptive, with as much attention as possible to previous knowledge, and thus to contribute to the mysterious process called insight. (p. 695)

In order for data to be "productively descriptive," the data must be understood. Fleming (1971), Goslin (1967), Mayo (1970), Brim, et al. (1964), Backman (1976), and Popham (1975) are only a few who have discussed the need for a greater understanding of basic data analysis and use of test data on the part of educators. In keeping with the perceived objectives of these authors, it is not the intent of this study to produce statisticians out of classroom teachers. However, there are some basic concepts of statistics and data analysis that appear to be essential for more effective understanding and use of test information.

These statistical areas are limited to measures of central tendency, mean, median, mode, standard deviation, shapes of distributions (skewness), standard error and correlation.

The rationale for the staff development program was based on simplicity of content covering the definition of test and measurement terms, statistics, and the application of acquired knowledge to local test interpretation and testing situations. An initial condition of program implementation was that the participants would also be the evaluators of the program. Both formal and informal feedback from the participants was used to revise the program after each administration in the pilot school districts. The formative process of feedback was a major factor in the development and evaluation of the program, and is discussed in more detail in the following chapter.

C H A P T E R I I I

DEVELOPMENT, IMPLEMENTATION AND EVALUATION OF A STAFF DEVELOPMENT PROGRAM ON TESTS AND MEASUREMENT

The model used in developing, implementing and evaluating the staff development program is shown in Figure 1. This model is an adaptation of the Provus Discrepancy Evaluation Model (Provus, 1971), which has four sequential stages: design, installation, process and product. The Provus Model, though defined as an evaluation model, places evaluation in the context of program development. Consequently, in the course of this study, the Provus Model is viewed as a combination development and evaluation model.

At each stage of the development/evaluation model, some indicator of performance is obtained and compared with a standard or criteria of performance. If a discrepancy is discovered, it is necessary to determine the reason for the problem and what actions are possible and most effective in correcting the situation. The necessary elements for program evaluation are: (a) criteria for identifying relevant evaluative information based on desired standards of performance, (b) new information about actual performance and

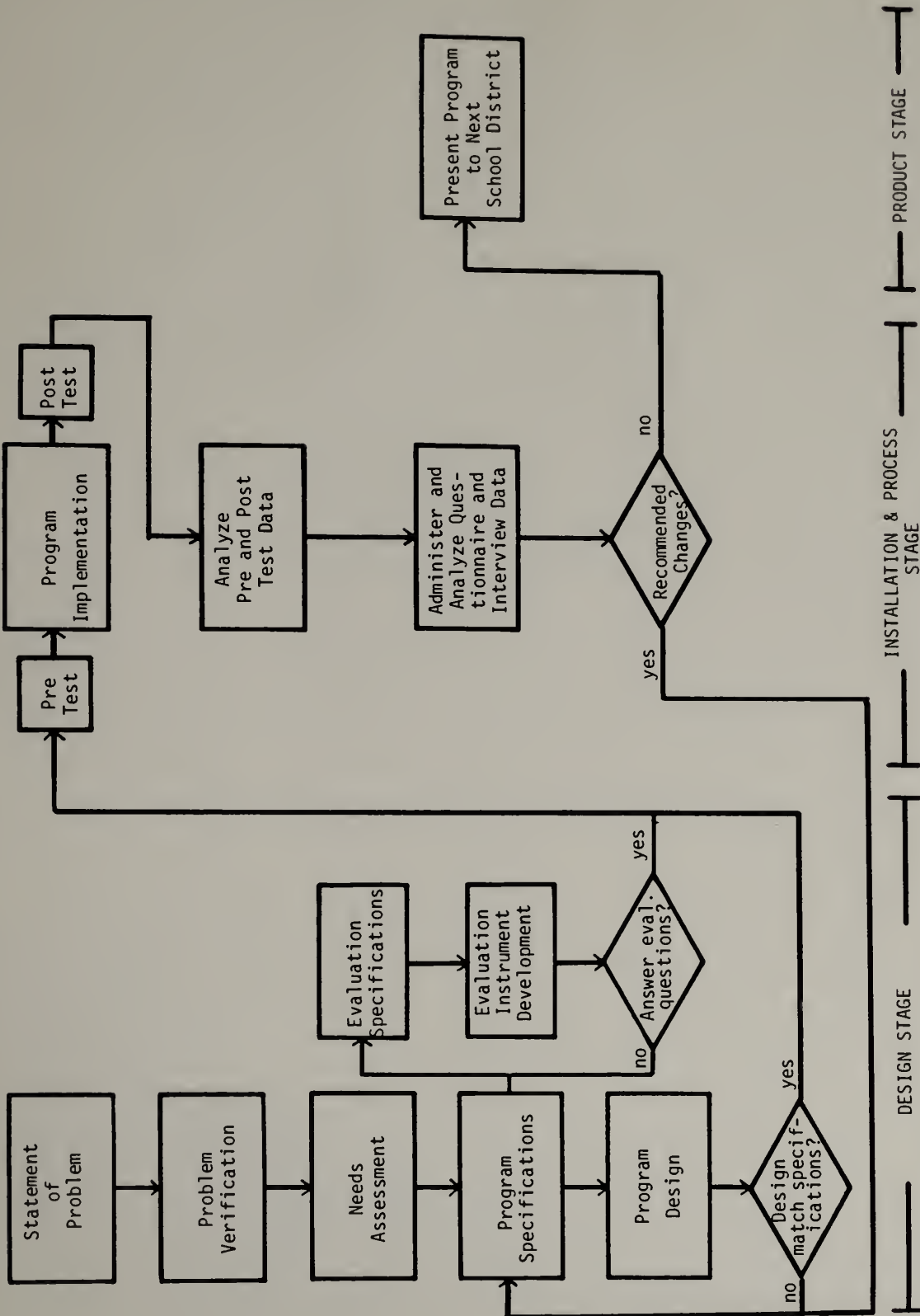


Figure 1. Model used in developing and evaluating the staff development program in tests and measurement.

practice, and (c) a decision to continue, change or terminate the project based on a comparison of information with criteria. Emphasis is placed not only on what should be produced (outcome), but also on how it should be produced (processes).

The basis of the Provus Discrepancy Evaluation Model is a formative evaluation process. The systems design methodology for the staff development program in tests and measurement also included a formative evaluation process. Scriven (1967), who originally defined the terms "formative" and "summative" evaluation, described formative evaluation as "process research, but it is of course simply outcome evaluation of an intermediate stage in the development of the teaching instrument." In contrast, summative evaluation is more like applied educational research. It involves the collection of data that will be used to determine "the effectiveness of the overall program, that is, the end product of the programs, not just the means of instruction, as in formative evaluation" (Asher, 1976, p. 205). Since this project is primarily concerned with building an instructional component, the emphasis in the developmental process is on formative evaluation. Three of the Provus stages in the development of the staff development program involve a formative evaluation process: Design, Installation and Process. The Product Stage involves summative evaluation in that the resulting program

is an end product of the program development cycle.

One reason the Provus Model serves effectively as both a developmental and an evaluative model is due to the iterative and sequential nature of its application. The iterative aspect occurs within each stage of the evaluation model. The results of each step within the model are reviewed by the evaluator and compared with the standards or specifications established for the particular stage. For example, the iterative aspect of the staff development program occurred in the Design Stage through continual review of the literature and subsequent modifications to the content specifications. Any discrepancies that were found were rectified prior to the implementation of the Installation and Process Stage. The iterative process was emphasized in the combination of the Installation and Process Stage. Continual formal and informal feedback from the participants was used to modify the program both during installation and prior to each presentation in a new school district. The sequential aspect of the staff development program occurred through the establishment of specific steps in the evaluation process, and the sequential development of the program through the systematic evaluation of these steps.

The specific developmental and evaluative steps outlined in the model shown in Figure 1 are described in the following sections.

Design Stage

Statement of Problem

The study was based on the problem described in Chapter I as a lack of skill on the part of educators in understanding and making effective instructional use of test information. This perceived lack of skill on the part of the educators appeared to be a genuine problem based upon the writer's experience over the past nine years in working with various Connecticut educational groups and was supported by a review of the literature. The major objective of the developmental activity was to create an effective program for upgrading the skills of the educators in making better instructional use of test data.

Problem Verification and Needs Assessment

A needs assessment is the process that establishes whether the situation being studied is actually functioning at a level below that which is expected. It answers the question of whether or not a problem really exists by showing if there is a discrepancy between what is and what should be. The primary sources of input to the present needs assessment were, (a) the literature, (b) external resources (e.g., college or university professors involved in teacher training), and (c) practicing teachers and administrators. The resulting information, as described in Chapter II, indicated that a need does exist for on-site

staff development programs in tests and measurement. There appeared to be a decided discrepancy between the skills possessed and those needed by educators in interpreting and using test data. The results of needs assessment should verify the problem statement and contribute to program specifications. Since the major input to needs assessment was a review of the literature, the assessment of needs served to structure the content and format of the material covered in program design.

Program Specifications

The program specifications were the product of the needs assessment. The program specifications led to the design of the program and met the standards for evaluating the design stage. The major variables to program specification and design were:

1. Target group. The primary program participants are classroom teachers and administrators since they are the instructional managers. Counselors are also considered an important audience for the training program for some expressed a need to upgrade their skills in tests and measurement. Counselors are also seen as a major contributor in the formative evaluation process, for most counselors have had more training in tests and measurement than the average classroom teacher. In addition, they have experience working directly with teachers and have an opportunity

to know the level of understanding about test results and the uses teachers tend to make of these results in the educational environment. Consequently, their background in tests and measurement plus their experience working with teachers gave them unique qualities as critics of the program.

An initial attempt was made to limit the number of participants to a maximum of 15 in the lecture part of the program and a maximum of 5 in the workshop activity. Small groups were desired to allow for the greatest possible interaction.

2. Time of presentation. The actual presentation of the program was designed to coincide as closely as possible with the period of the year that testing was scheduled. The time involved in administering the program was originally established at 6 hours, including a basic workshop activity. It was understood that local time constraints or training needs could shrink or expand the original 6 hour time specification. In keeping with the research by Hastings, et al. (1961), the program was designed to be divided into at least three sections that could be presented over a period of time -- preferably a week or two apart. The specifications, however, called for a variety of timing alternatives from a single day presentation to a time span of a month between sessions.

3. Location of presentation. The location of the staff development program was planned to be on-site. The program specifications called for a flexible training vehicle which could be easily transported to the school community desiring training. For purposes of evaluation, the original plans specified the implementation of the staff development program in at least five school districts.

4. Mode of presentation. Since the review of the literature indicated that many professional educators lack a knowledge base in tests and measurement, the primary mode of presentation was a lecture format. About 70 percent of the staff development program was specified to be in a lecture format with the remaining part designed as a "hands-on" workshop where participants could experience working with their own data. The workshop activity could be accomplished with the entire group or the groups could be divided into smaller units.

5. Materials. Materials consisted of both overhead transparencies for the instructor's use and hard copies of these transparencies to be provided to the participants in a reference notebook.

6. Content specifications. The review of the literature emphasized certain major content areas that were characteristic of existing texts and courses on tests and

measurement. These materials focused on, (a) the description of terms -- methods of describing the instructional environment through testing, (b) basic statistical techniques -- statistical techniques for highlighting potential problem areas, and (c) application -- improving instruction through understanding and applying data. Consequently, these general content areas served as the basis for establishing the initial content specifications.

The staff development program was divided into three parts. These three parts and the specific content specifications are described below. The item numbers of the pre- and post-assessment instrument (see Appendix A) that were used to measure the program's effectiveness are shown in parentheses following the description of each major heading.

Part I. Methods of describing the instructional environment through testing.

Definition of terms

Measurement

Testing

Assessment

Evaluation

Formative

Summative

Norm-referenced tests

Purposes

Influencing variables

Examples

Criterion-referenced tests

Purposes

Examples

Differences between Norm-referenced and
Criterion-referenced tests
Purposes of each

Teacher-made tests
Purposes
Construction
Analysis
Interpretation
Application

Part II. Statistical techniques for highlighting
potential problem areas.

Basic statistics (2, 5, 14)

Mean
Median
Mode
Standard deviation
Correlation
Standard error
Measurement
Mean
Confidence Intervals

Shapes of distributions (4, 5, 9)

Normal
Skewed
Compressed
Bi-modal
Relation to frequency distributions
Instructional applications

Reliability (2, 13)

Definition
Application

Validity (12)

Content
Criterion-related
Construct

Part III. Application -- Improving instruction
through understanding and applying data.

Types of scores (4, 11, 14)

Raw Scores
Grade Equivalent Scores
Standard Scores

Definition
Derivation
Application

Norms (1, 10, 11)
National percentiles
Local percentiles

Public display of data (4, 9)
Narrative reports
Graphic displays

Application of local data (3, 6, 7, 8,
14, 19)
Interpretation of individual reports
Interpretation of group data reports

The section on "Types of Scores" in Part III was considered an important section in this phase of the training program. Particular emphasis was placed on the use of grade equivalent scores. The literature continues to discuss the disadvantages of grade equivalent scores and this topic is occurring more frequently in the professional literature. However, test publishers continue to advertise grade equivalent scores in their documentation and even use them in the examples of their reports. Consequently, the credibility of these types of data is fostered and amplified by the fact that most educators were trained to use them.

Standard scores were stressed as an alternative to grade equivalent scores. Since standard scores are based on an equal interval scale, they do have some distinct statistical advantages over grade equivalent scores. These equal interval scaling properties of standard scores also

provide the potential for more effective plotting of growth characteristics and out-of-level testing. The original program design did not provide for much detail on the derivation or use of standard scores. The process of continual review and modification, provided through the development and evaluation model, allowed for revision of the section on standard scores to meet what appeared to be a weakness in the literature and a stated need of the participants.

7. Evaluation specifications and design. The evaluation strategy was intended to be a formative process with a major emphasis on feedback from program participants. Each stage of the evaluation process had specific questions which covered the most significant aspects of the respective stages. The specifications for all the evaluation instruments were a result of these specific questions raised in each stage of the development/evaluation mode. Evaluation specifications also included provisions for informal review of the program through both oral and written responses from the participants during and after each presentation.

Formal instruments for program review and modification were designed to assess the value of the content specifications and provide feedback on the overall effectiveness of the program. These instruments consisted of

pre- and post-assessments of the training programs impact on cognitive skills and attitudes in tests and measurements (see Appendix A). A questionnaire (see Appendix B) and an interview form (see Appendix C) were designed to provide feedback for program modification and improvement. These instruments were implemented in the Installation and Process Stage and will be discussed in more detail in that section.

Program Design

All of the documentation cited in the reference section and reviewed in Chapters I and II was examined for examples of information which could be included in the staff development program. The intent was to use either material that existed or that could be modified to fit the content specifications. If material did exist that was amenable to the design, then permission to use it was sought from the authors or publishers. If material was not found or not appropriate, it was developed or modified with the permission of the original author. Each aspect of the design phase was viewed against the program specifications to insure that the standards were being considered.

An original set of 75 overhead transparencies was developed according to the overall content and sequence of material outlined in the program specifications. Copies of these transparencies were also provided in a notebook for

the participants to use as future reference material. These copies did not contain narrative information. The narrative explanatory material was added as the program was implemented and comments were received on the oral presentation. The oral presentation was a variable in that different approaches were used in discussing the content in order to establish the most effective descriptive material. The evaluation feedback was used to help in structuring the content of the narrative material. Final narrative was written for each transparency during the month of December, 1976. The narrative description along with the transparencies was given to a reporter on the staff of the Hartford Courant, a daily newspaper serving the State of Connecticut. This reporter is noted for his clear and objective reporting, and is not knowledgeable in the area of tests and measurement. This review procedure was used to ensure that the narrative material was as free as possible of any terminology that would be confusing to the layman and that it conveyed, clearly and simply, the information on the transparencies. Though the narrative material was not a part of the formal evaluation process, the input to it was a result of the comments received from the questionnaire, interviews and informal feedback from the participants.

Evaluation of the Design Stage

The design of the program involved establishing and verifying that a need for training existed, determining specifications to satisfy the need, and designing a program which reflected the stated specifications. Evaluating the Design Stage involved comparing the design of the program against a set of design criteria. The design criteria included the following program specifications: target group, time of program implementation, location of program presentation, mode of presentation, types of materials used, and content specifications. The major design criteria in the Design Stage were the content specifications. The evaluation process in this stage focused on insuring that the design criteria, or specifications, were adequately defined and that the program specifications were complete. This was accomplished through a continual review of the literature and the variety of existing resources for confirmation or restatement of needs or content specifications. In addition to the process of continual review and modification, the design also included an iterative process of program implementation, review and modification. The basic set of program specifications was constantly subjected to review and modification as a result of the program being implemented and evaluated in five school districts.

The major question in the Design Stage was determining whether or not the program specifications were complete.

However, the iterative process of program implementation, review and modification also helped in defining the needs of educators, further clarifying program content, and establishing the basis for responding to the following questions:

1. Is there congruence between the information test publishers think is essential and that perceived necessary by school personnel for effective decision-making purposes?
2. What type of information do teachers and administrators indicate they need to have in order to make their use of test data more effective toward improving instructional practices?

Installation and Process Stage

As stated earlier, due to the non-repetitive nature of program implementation in any one school district, the Installation and Process Stage were combined in this study. The combination of these stages seemed particularly practical due to the single implementation of the program in each district, which resulted in a limited time involved in data gathering, and the necessity of immediate application of these data to program modification. The data gathering process required the development of evaluation instruments and techniques which could obtain as much information as possible in a relatively short time span. The time specified for program implementation would not allow for more than a total of 30 minutes for both pre- and

post-assessment.

Program Implementation

Schedules for program implementation were established in September, 1976. Six school districts were contacted and asked to participate in the staff development program. All six districts agreed not only to participate, but to engage in the formative evaluation process. Table 1 shows the school districts participating, the number of participants (first session) and the total time spent in actual program activity (not including pre- and post-assessment). In all but one case, the size of the groups exceeded the desired number of participants mentioned in the program specifications. The average time spent per district was three hours and fifty-one minutes.

Due to local needs, in most districts the program was divided into two sessions with varying time spans between sessions. There were a total of 134 educators present at the first sessions, exclusive of Glastonbury, and 120 present during the second sessions. Fifty percent of this loss between sessions was due to an out-of-town conference called in one district too late to change the training schedule. Two of the five districts requested follow-up activities related to more intensive interpretation of actual school results. A total of 12 teachers and administrators were involved in these follow-up activities and

TABLE 1

School Districts Participating in the
Staff Development Program on
Tests and Measurement

District	<u>n</u>	Number of Sessions	Time Between Sessions	Actual (total) Program Delivery Time
Pittsfield, MA	27	2	1 Month	4 Hrs. 40 Min.
Enfield, CT	48 ^a	2	1 Week	3 Hrs. 25 Min.
West Hartford, CT	14	2	2 Weeks	3 Hrs. 55 Min.
Newington, CT	21	1	-	4 Hrs. 15 Min.
Farmington, CT	24	2	1 Week	3 Hrs. 35 Min.
Glastonbury, CT ^b	18	2	1 Week	3 Hrs. 30 Min.

Note: In order to preserve district anonymity, the school districts are not listed in the order of program implementation.

^a Enfield was divided into two groups. Each group received approximately the same amount of training.

^b Glastonbury participated in the final version of the program. They did not participate in the same evaluation process as the other districts.

each received approximately 45 minutes of instruction. The major topics discussed were the interpretation of individual data reports for instructional purposes and the reporting of data to parents. At the follow-up sessions, the primary materials used were individual pupil test records and group item analysis reports.

The composition of the groups differed markedly. The Pittsfield, Farmington and West Hartford groups were primarily composed of teachers. Enfield had two groups; one was primarily composed of administrators and the other was composed of teachers and reading specialists. The Newington group consisted of administrators. Glastonbury, the group to experience the completed program, was composed of teachers, administrators, counselors and educational specialists.

Vital to the formative evaluation process is the quantity and quality of feedback. If the formative evaluation process is to be used in the development of any program, it is imperative that sufficient allowance be made for feedback from the beginning of the design through the product stage. The evaluation of this staff development program concentrated on this process of feedback. The effectiveness of the process was highly dependent on the cooperation received from local school administrators in supporting the program implementation and on the active interest and involvement of the participants. Prior to program implementation in each district, time was devoted to

explaining the developmental status of the program, emphasizing that the audience was to play a participant role in the shaping of its final form. The processes involved in program implementation and information gathering were explained in detail before the program was implemented.

Anonymity was preserved on all responses except the interviews that were held with a random sample of the participants. However, those interviewed were told that their individual responses would not be identified in any way.

Evaluation of the Installation and Process Stage

In the Provus Model, the Installation Stage involves a comparison between the results desired from installing a program and the actual results obtained. The Process Stage involves gathering information from all variables which can influence the design, development or implementation of the program rather than the end product. The Process Stage is concerned more with enabling than with terminal objectives.

In the staff development program, information gained through the Process Stage of evaluation was both a result of and had a direct effect on the activities in the Installation Stage. The pre- and post-assessment instruments and the questionnaire and interview procedures were designed to elicit discrepancies between desired and actual results.

The evaluation questions addressed in this stage dealt with three major areas. The three areas are:

(1) Knowledge -- the degree to which the participants improve their cognitive skills in tests and measurement, (2) Attitudes -- the effect of the program on the educators' attitudes about tests and measurement, and (3) Participant reaction -- the extent to which the program participants contribute toward and are satisfied with the content and process of program implementation.

1. Will a staff development program on tests and measurement contribute significantly toward a teacher's or administrator's knowledge and use of test data?
 - a. Pre- and post-assessment of cognitive skill data.
 - b. Interview data.

2. What attitudes do teachers and administrators have about tests and measurement as used in the public school environment before and after implementation of the staff development program?
 - a. Pre- and post-assessment of affective data.
 - b. Questionnaire and interview data.

3. Do the participants in the staff development program indicate they are satisfied with the content and process of program implementation?
 - a. Questionnaire and interview data.
 - b. Informal feedback.

Instrument Design and Development

The pre- and post-assessment measures of teacher attitudes and cognitive skills on tests and measurement were

adapted from an instrument used in the Hastings, et al. (1961) study. The Hastings instrument was a multiple choice test entitled Knowledge and Interpretation of Tests (KIT). The KIT Test consisted of 60 items with reliability coefficients ranging from .68 to .74. Seventy-three percent of the items on the KIT Test had biserial coefficients with the total test score of .30 and above. No validity data was reported. The instrument adapted for this study (see Appendix A) was administered both before and after program implementation. The instrument had questions which were directly related to key content specifications in addition to questions gathering information on general background characteristics. As a result, it was designed and used more as a criterion-referenced instrument than the earlier Hastings KIT Test. The reliability¹ of the adapted pre- and post-assessment instrument ranged between .55 and .63 with 73 percent of the 11 items having biserial coefficients with the total score of .30 and above. The relatively low reliability coefficients are probably due to the small number of test items, heterogeneity of the items, and the relatively homogeneous sample resulting in a restricted range of potential variability. Though total scores were used in interpreting the results,

¹Spearman-Brown formula for estimating reliability from average item-test correlations.

greater emphasis was placed on changes in the terms that related to specific content areas. The content and question numbers on this instrument are shown in Table 2.

Since there was a desire to match the participants' pre-test and post-test responses, each individual was presented with a small manila envelope containing a 3 by 5 card with a unique number written on the card in pale yellow ink. The numbers on all cards were randomly selected from 1 to 300 with no duplicates. The participants were asked to write the number they found on the card inside their envelope on the top of their test sheet, place the card back into the envelope, seal it, and write their names on the outside of the envelope. The participants were told they would be taking a similar test again at the completion of the program. At that time, their envelopes would be returned and they would be asked to open them and record the same number on the top of their post-test sheet. Both the cards and the envelopes would then be their property to dispose of at their discretion. Since the envelopes were of heavy buff stock and the numbers on the cards inside were written in pale yellow, it was obvious that any attempt to read the number through the envelope would be futile. It was deemed desirable to remove as much threat as possible from the testing situation and this procedure was expressed by some afterwards as being a positive approach in this direction. The pre- and post-assessment

TABLE 2
 Content Categories and Item Numbers
 Included on the Pre/Post Assessment
 Instrument

Content Categories	Item Numbers						
Background Information	21	22					
Attitudes Toward Teaching	3	6	7	15	16	19	
Knowledge and Use of Tests, Scores and Norms	1*	8*	10*	11*	14*	17	20
Reliability and Validity	2*	12*	13*				
Distribution Characteristics	4*	5*	9*				

Note: See Appendix A for example of the instrument. Item 18 is a general information item and does not relate to any of the above categories.

* Items scored as part of the cognitive skills pre- and post-test (11 items).

information was used to determine discrepancies between the content transmitted and the skills acquired as measured by these instruments.

In addition, an anonymous questionnaire (see Appendix B) was developed to be used with all participants at the conclusion of the training to get formal feedback on how well the program met their particular needs. This questionnaire was designed to provide feedback on various aspects of program design, installation and process.

An interview form (see Appendix C) was also designed for the purpose of gathering more detailed and personal information from the participants. The interview process was conducted at the conclusion of the training on a random sample of 79 of the participants.

Product Stage

The Product Stage of the program development and evaluation cycle answers the question of whether or not the end product or program is successful. The terminal objective of the staff development program was to provide educators with test and measurement skills that would enable them to make more effective instructional use of test data. Questions addressed in this stage dealt with the participants' use of the information they learned through the program and the extent to which they used this information to improve instructional practices. Specific questions responded to

in the Product Stage were:

1. Does the basic statistical portion of the staff development program improve the interpretation and use of test data?

Questionnaire and interview procedures.

2. Has the approach toward more effective instructional use of test data brought about any changes in the instructional application of test information?

Interview procedures.

C H A P T E R I V

RESULTS

The original staff development program on tests and measurement was designed during the summer of 1976. The content and format was based on a review of material covered in commercial texts and material available through courses taught at teacher training organizations. The evaluation design called for the implementation of the program in at least five school districts with a variety of demographic characteristics, varying target groups and implementation times. The population characteristics included urban, rural and suburban school districts. The program implementation time ranged from a single one-day session to three sessions given over a three-month time period. The participants included an approximately equal mix of teachers, administrators, and special education personnel with the addition of some guidance counselors. A prerequisite to program implementation in all districts was the agreement of the participants to become actively involved in the evaluation process. This included their performance on a pre- and post-assessment instrument, response to a questionnaire, participation in an interview process if selected, and the

provision of informal oral and written comments during program implementation regarding the content or format of the presentation.

This chapter presents an overview of the program development activities and the results of the summative and formative evaluation procedures. Both the summative and formative evaluation results are presented concurrently as they relate to the specific stages of the evaluation model.

Evaluation Processes and Results

Design Stage

The format and content of the staff development program followed that found in most basic texts and reflected the needs of educators as evidenced in the review of the literature. The major evaluation process in the design stage involved the continual comparison of program content with material available in recent texts. In addition, the results of recent studies on the use of tests and measurement were examined for evidence of training needs as stated by educators. The staff development program was modified to address any need discovered which did not have content coverage. The feedback procedures of questionnaire, interviews and informal comments indicated that the participants in the staff development program were generally satisfied with the program content and approach. There were also some comments which indicated that the educators liked this

"participant evaluator" approach to training. It was stressed from the beginning that they should be critical listeners. They seemed to appreciate the fact that they too were providing information -- that this was a two-way process.

Evaluation of the Design Stage. The three basic questions addressed in evaluating the Design Stage dealt with (a) the completeness of the program specifications, (b) the congruence between information test publishers provide and that desired by educators, and (c) the types of test information needed by educators.

Completeness of program specifications. The primary concern in the evaluation of the Design Stage was insuring that the program design was a result of the program specifications, and that the program specifications were complete. The most important consideration in this matching process was the program content. The original content specifications were a product of the review of the literature. The program was designed based on these specifications. Both program specifications and program content were modified as new resources were examined. This process of literature review, content specification development, and program design was an on-going activity from January through August of 1976. The result was a basic program designed with specific content, materials and implementation procedures.

Further changes to the program specifications, particularly the program content, were a result of participant feedback during the Installation and Process Stage. There were changes made in the program specifications as a result of the participant feedback. The specific changes are presented in this chapter in the evaluation of the Installation and Process Stage.

Congruence between information test publishers provide and that desired by educators. In response to the question regarding whether there was a match between the information provided by test publishers and that desired by educators, there appeared to be some discrepancy. The majority of educators appear to have a very limited knowledge base about what is available with respect to test results. When they are exposed to such information as graphic frequency distributions and item analysis data showing skill deficiencies in a simple and concise manner, these become desired needs that many test publishers do not provide. Questionnaire and interview responses indicated that educators would have made more use of test information if they had known certain types of information were available. The staff development program content includes examples of test results from test publishers who seem to make the most effective use of data -- particularly in the presentation of information. For example, item analysis is presented very differently depending on

which scoring service the school is using. When educators see examples of the various ways these data can be presented, they begin to anticipate different ways these data can be used. The encouragement of an informal feedback process during program implementation stimulated considerable discussion about the various types of information that were available.

Another example of a type of report few test publishers of norm-referenced standardized achievement tests provide is a report which groups pupils on the basis of tested skill deficiencies. An extension of this type of report is a narrative report which one test publisher provides that gives a detailed description of skill deficiencies by school and by district. Educators need to be shown examples of what is available, particularly since the computer has greatly increased the potential for information processing and reporting.

It is also important for educators to see the difference between norm-referenced and criterion-referenced test results. Many of the educators in the staff development program had their first experience seeing examples of commercially produced criterion-referenced test reports. These types of reports provide an entirely new outlook on testing and also emphasize the potential of item/skill response data rather than the reliance on scores as the only indices of pupil progress. Discussion on the uses of reports also

provided useful information as to how current information or formats could be made more effective. Twenty percent of the participants, primarily from one district, found that certain tests they had been using for years were not providing the type of information they thought they were getting. They reported a greater desire to focus first on their specific needs for testing, then identify instruments or techniques to meet those needs.

Types of test information needed by educators. The information obtained in the previous section is also applicable to the question of the types of information desired by educators. Forty-eight percent of those interviewed stated they were not satisfied with the kind of testing being done in their district, and 57% stated a greater desire for more criterion-referenced test information (see Appendix E). The majority, 84%, approved of achievement testing as a way of helping them identify skill deficiencies in their pupils. The desire appeared not to be against testing, but for norm-referenced tests which could offer more diagnostic information. The traditional test information received in the districts in this sample were test scores, with little or no availability of item analysis information.

As reported in the review of the literature, studies have shown that educators are not generally dissatisfied with tests or testing, but they do feel inadequately trained in using the results. In this study, 78% of the

questionnaire respondents felt that achievement testing was a worthwhile activity and the same percentage claimed the staff development program helped them to be more realistic in their use of test results (see Appendix F). It appears that there are two basic test information needs for educators: (1) practical in-service training on how to use the results they currently receive, and (2) more emphasis on item analysis information presented in format that is easy to interpret and use.

Installation and Process Stage

There were three key questions addressed in the evaluation of the Installation and Process Stage. The three questions dealt with (a) the effectiveness of the program in increasing the knowledge base of the participants in tests and measurement, (b) the attitudes of the participants toward the area of tests and measurement, and (c) the program participants' reaction to the content and process of program implementation.

The analysis of the data shows results from six school districts; however, there were only five school districts in the study. The reason for this apparent discrepancy was that one school district was divided into two groups, and each group is displayed as a separate "district."

Improvement in cognitive skills in tests and measurement. Table 3 shows the proportion of correct responses in

TABLE 3

Percent of Correct Responses in Each of the
Cognitive Skill Areas as Measured by the
Pre- and Post-Assessment Instrument

District ^a	Knowledge and Use			Reliability and Validity			Distribution Characteristics		
	Pre Test	Post Test	Diff.	Pre Test	Post Test	Diff.	Pre Test	Post Test	Diff.
A	52	60	+ 8	64	77	+13	68	67	- 1
B	54	61	+ 7	77	80	+ 3	53	73	+20
C	62	64	+ 2	76	82	+ 6	45	70	+25
D	48	60	+12	71	84	+13	57	73	+16
E	46	58	+12	63	73	+10	38	60	+22
F	48	67	+19	63	84	+21	56	84	+28
Total	51	61	+10	68	80	+12	54	71	+17

^adistricts are shown in order of program implementation.

each of the cognitive skill areas as measured by the pre- and post-assessment instrument. The districts are listed in the order of program implementation. There was a general increase in the percentage of correct responses in the three cognitive skill areas in all districts, except District A. The participants in District A were not providing positive feedback on the section dealing with distribution characteristics. The transparencies dealing with this content area were changed prior to program implementation in District B. The results, as shown in Table 3, were far more positive.

The general pattern of improved performance from pre- to post-assessment coincided with major program modifications. For example, major changes were made to the program in the area of interpreting norms (percentiles), the derivation and meaning of standard deviation, and the meaning of reliability -- particularly as it relates to standard error. These changes were introduced into the program prior to implementation in District D. There is also a pattern of the percentage of correct responses to increase as the program continues through the iterative process of implementation and modification.

Table 4 shows the changes in total test scores from pre- to post-assessment. Significant differences between the pre- and post-assessment total mean scores were achieved in all but two districts. The analysis of variance F values

TABLE 4

Pre- and Post-Assessment Results on Cognitive Skill
Information About Tests and Measurement

District	Pre-Test		Post-Test		Difference		
	Mean	SD	Mean	SD	Between Means	t	r
A	6.6	1.7	7.3	1.7	.7	1.739	.16
B	6.6	1.6	7.7	1.3	1.1	2.304*	.03
C	6.7	1.4	7.7	1.3	1.0	1.618	-.16
D	6.2	1.3	7.6	1.6	1.4	3.874**	.49*
E	5.4	1.5	6.9	2.1	1.5	3.204**	.31
F	6.0	1.2	7.9	1.1	1.9	7.869**	.54*
Total	6.2	1.5	7.5	1.6	1.3	7.114**	.25*
Differences Between Districts	F = 2.28 $\bar{d}f = 5,106$		F = 1.06 $\bar{d}f = 5,106$				

Note: Maximum score possible = 11. Only those participants with complete data were used in the above analyses -- those who took both the pre- and the post-test ($n = 112$). Districts are listed in order of program implementation.

* $p < .05$
** $p < .01$

at the bottom of the table indicate there are no significant differences between the districts on the pre-test or the post-test means. The majority of non-significant correlation coefficients between pre- and post-test results indicate that, with the exception of two districts, even though the scores tend to show an overall increase, the increases reflect some lack of consistency with respect to the growth within most of the districts. The two groups that have the significant correlation coefficients between the pre- and post-test results are composed almost entirely of administrators.

Attitude assessment. The pre- and post-attitudinal information (see Appendix D) revealed a 10% increase in those who disagreed that objective measures may have a negative effect on learning. The group maintained their relative position in believing that group data had little value in instructional planning for individual children. There was only a 2% change favoring those who felt group data had individual instructional relevance. Post-assessment results revealed a 4% increase in those whose general opinions about the use of test results had improved. There was also a shift from 38 to 46 in the percent of those believing it was a good idea to have a yearly testing program of abilities or aptitudes, and the relative percent of those in favor of yearly achievement testing

remained at 86%. Eight percent of the participants changed from a pre-test attitude that tests answer many questions they have about students to a more general opinion that test data is a valuable piece of information useful in raising important questions in their minds about students.

Interview data (see Appendix E) were highly supportive of testing, particularly in the use of test data to help in identifying skill deficiencies. However, 57% did report that they would like to see criterion-referenced tests replace norm-referenced tests.

The quantitative responses to the questionnaire administered after the program (see Appendix F) indicate a generally favorable attitude toward testing. Approximately 70% of those responding to the questionnaire made requests for more use of criterion-referenced measurement in their schools.

The staff development program did not appear to greatly improve testing attitudes, for they were at a relatively high initial level. However, the program has appeared to place the whole issue of tests and measurement in a different, if not better, perspective. Part of this may be due to the general low level of initial knowledge about tests and measurement. Fifty-seven percent reported either none or one course in tests and measurement in college. Fifty-five percent indicated no training in tests and measurement since college.

Participant reaction to program content and implementation. Out of the 134 educators present during at least one session of the program, 58% responded to the questionnaires which were returned through the mail, 59% submitted informal written comments regarding the program's strengths and weaknesses, and all of those randomly selected to be interviewed participated in the interview process. In four out of the five districts, at least 40% or more of the participants offered oral comments either during or after the presentation. In all cases, the groups were attentive and actively participated in the formative evaluation process. There were many oral as well as written comments after each presentation in addition to the questionnaire and interview responses collected later in the school year. All of these feedback mechanisms served as a basis for comparing the actual implementation practice to good generic standards of program implementation -- making certain that the format and material was easy to follow and understand. Recommended changes were made in the program or method of delivery prior to the next implementation.

There were two sources of written comments from program participants. The questionnaires provided for written comments to specific questions about the program. In addition, participants were encouraged to respond informally, either in writing or orally, regarding general concerns about the program. Six percent of the questionnaires'

written responses reflected negative attitudes toward the program in that participants did not feel it was appropriate for them. Seven percent of the informal written responses expressed no complimentary comments, but did offer suggestions for program improvement. All the rest were favorable with constructive criticism about specific areas that could be strengthened.

Some problems with both content and format were discovered early in program implementation. Specific changes made in the program were based primarily on participant feedback and are shown in Appendix G.

Weaknesses were revealed early in the program and centered on the complexity of the statistical language, technical aspects of data interpretation, desire for more emphasis on the application of acquired knowledge, and more time (sessions) to allow for a slower pace in presentation. Immediate attention was given to simplifying the language and technical aspects of the program. It was agreed by most of those involved in the one-day session that this time span was too compressed for adequate comprehension of the material. Objections were raised over several of the overhead transparencies in the statistical section and particularly the transparency showing the derivation of standard scores. These and other prerequisite transparencies were modified and expanded through subsequent presentations until the comments from participants reflected

satisfaction with the material. The consistent objection to speed of delivery was a difficult problem to correct due to the time limitations in most locations. This proved to be a realistic operational constraint which in itself caused program modification. The iterative aspect of the formative evaluation process provided an additional benefit in that the program objective of flexibility was easily achieved. All questionnaires and written comments were examined immediately after each presentation for any indication of a problem area or lack of clarification. Recommended revisions were made prior to the next presentation.

There were several specific weaknesses in the content which needed to be corrected in order to meet the overall program objective. The definitions for the terms Measurement, Testing, Assessment and Evaluation used in the initial program were not clear and distinctive enough. The interview sessions revealed that these terms needed to be defined, but some of the participants were still vague about relative distinctiveness after experiencing the program. Consequently, the definitions were changed for greater clarity and the term Evaluation was expanded to include a discussion of the formative and summative process.

Part I of the program had no initial provision for examples of information returned to teachers from norm-referenced and criterion-referenced measurement devices. This was stated by participants as being a weakness which,

if rectified, would contribute toward clarifying the differences in their respective uses and interpretation. Therefore, examples of both types of these measurement vehicles were included and emphasis was placed on how they could offer different and complimentary information.

An early noted omission was the absence of any reference to teacher-made tests. Since so much emphasis in the program was being placed on commercial instruments, some participants were getting the impression that the commercial route was being preferred to any other method of testing. A section on teacher-made tests was included in Part I of the program with emphasis on the need for teachers to state a purpose for the test, construct an instrument to accomplish that purpose, analyze results to determine the instrument's instructional validity, and apply the results in an instructionally efficient manner.

The definitions for the different types of validity were criticized in that they were in question format. For example: Content Validity was defined by the question: Does the test measure the sequence or types of skills covered in the curriculum? These terms were changed and the content revised to reflect the more recent and accepted types of validity and the definitions were amplified in narrative format. This approach received more favorable response, particularly since the definition was more useful as a reference source.

A major problem expressed by participants was in understanding standard scores. Since correlation and reliability are important prerequisite skills for understanding standard scores, these two areas were moved to an earlier section of the program. Many of the problems in understanding standard scores seemed to diminish with the movement of correlation and reliability to Part II of the program. The section on "Types of Scores" was originally in Part II and was moved to Part III, since it seemed to give added strength to the application section and some of the participants expressed a desire for this move.

The major modifications in the statistical section dealt with the subjects of standard deviation and the normal curve. The derivation of standard deviation was originally presented with one transparency but the concept was not grasped satisfactorily. Consequently, standard deviation was amplified into greater detail and made the subject of three transparencies which resulted in more favorable comments. The example of the normal curve was not receiving the type of favorable response desired until a grade equivalent scale from a norm-referenced standardized achievement test was added. The inclusion of a grade equivalent score scale gave it the "universal language" appeal and also dramatically revealed the unequal interval properties of that particular scale. This feature helped considerably in future discussions of the advantages and disadvantages of

different types of score scales.

The section on standard error in Part III was criticized as not being easily understood. Consequently, a transparency was added which showed the relationship between standard error, standard deviation, reliability, and the probability of a pupil's score being in a particular range. The emphasis was changed from the derivation of standard error to understanding and applying the concept of standard error to interpreting a pupil's score. This approach received more favorable responses and was amplified in subsequent sessions involving the interpretation and application of individual pupil data. Some of the specialists interviewed claimed they were pleased to have the overview on standard error. They stated that this would definitely influence the way they would interpret test data in the future and would cause them to be less dependent on a single score as a primary basis for decision-making.

Procedures and examples of displaying group data were greatly modified during program implementation as a result of participant feedback. The major direction was to emphasize a graphic approach to presenting group data rather than a numerical tabular procedure. Also emphasized were the various ways both norm-referenced and criterion-referenced group data could be displayed. A standard training package for Part III evolved from the expressed needs of the participants. It was in this section that the emphasis

was placed on the application of group data.

The participants voiced strengths both in the program's content and process of implementation. Content strengths were more noticeable in the latter versions of the program. The information obtained from participants through the interviews and the questionnaire indicated that the strengths of the program were primarily in the simplified handling of statistics and scores, making educators more sensitive to the problems associated with data interpretation, exposing educators to different types and ways of presenting and using test information, and general favorable remarks about the scope and sequence of the material presented.

Most of the strengths in the content areas were reported as dealing with the information in the statistical section, the section on types of scores, and the exposure to different types of tests and scoring formats. Most of the educators were not aware of the different types of tests that are available, what the tests are designed to accomplish, and the different ways computer technology has helped in displaying test information to make it more instructionally useful. Strengths in implementing the program were reported as being in the scope and sequence of the material presented. There was considerable satisfaction expressed over the way the program was structured -- typical comments included: well organized, complete, comprehensible,

quick moving, and interesting. There were no negative comments about the way the material was presented except in the initial presentation when objections were raised over the use of both transparencies and hard copy. Several participants stated that they thought this was a demeaning practice since the same material was presented in two different ways. Subsequent administrations of the program involved continued use of handouts. However, before the handouts were distributed, emphasis was placed on their use for supplemental notetaking and as reference material. When this use was stressed, the negative reactions ceased. All subsequent presentation received praise over this procedure -- in fact, this was a strong point expressed about the program.

In all districts there was ample support on the part of the local school administrators and active involvement on the part of the participants. The participants were encouraged to be critical of the program -- particularly any part of the content they felt was not clear or poorly presented. Both the quantity and the quality of the responses, written and oral, seemed to give support to the efficacy of the formative process as a valuable means of program development and evaluation.

Product Stage

There were two questions related to the Product Stage. These questions dealt with the end product, or the developed

program, and were concerned with (a) the value of the statistical portion of the program on improving the educator's interpretation and use of test data, and (b) the impact the program had on the instructional application of test information.

Knowledge of statistics in the use of test data. The statistical portion was considered a significant part of the staff development program. Statistics is one area the literature confirms to be a weakness with most educators. A basic knowledge of statistics is also important in order to understand both the strengths and limitations of test data.

Most of the participants responded favorably to the material on statistics. The greatest objection to this section was not about the content, but was in reference to the speed with which it was covered. Several expressed the desire to spend more time on this section, particularly that part which dealt with the derivation of scores. The questionnaire results (see Appendix F) indicated that 82% responded affirmatively that the program gave them a better understanding of the basic statistics used in testing. The interview data (see Appendix E) supported this information by confirming that the statistical section was largely responsible for the 78% who stated the program helped them to be more realistic in their use of test results. Thirty-three percent of those interviewed claimed the statistical

portion was too complex or involved for their needs. However, over half of the 33% came from the first district receiving the program. They contributed a large number of suggestions for the improvement of the statistical section. These improvements were incorporated into the subsequent sessions which resulted in a definite decrease of negative responses.

Narrative responses from the questionnaire were very supportive of the statistical part of the program. For example, teachers reported they felt more comfortable in dealing with parents after having a better understanding of standard error and how it relates to confidence intervals.

Instructional application of test information. Most of the participants were very positive regarding the program's contribution to their understanding and instructional use of individual data. Many comments both from the questionnaire and the interviews reflected appreciation for a greater awareness of the strengths and weaknesses of testing. The primary areas of greatest expressed interest for the application of individual data were the program's emphasis on percentile bands (and how they were derived) for use with parents, and the use of item analysis data for more definitive objective information. They saw both as a means of dealing with parents on a more informative basis, helping to identify skill deficiencies, and establishing a

basis for correcting skill weaknesses.

Administrators viewed the program as something they wanted their teachers to experience. During the course of the school year, they saw evidence of teachers using the item analysis information in identifying skill deficiencies, instructional grouping, and as an aid in articulating instructional material and emphasis from elementary to junior high. However, the use of item analysis data for grouping purposes rated low on the list of priority activities. Administrators also reported a more comfortable feeling of being able to explain data to parent groups, thus providing a better basis for public relations.

One comment which was relatively prominent was that teachers felt they did not need test data to identify skill deficiencies in their pupils. Given a normal classroom situation, it did not take them long to determine which pupils had problems. However, test data of a more diagnostic nature was perceived as a tangible and objective means of supporting their findings. Both teachers and administrators expressed this as a positive aspect of being exposed to these types of available information.

Seventeen percent of the participants responding to the questionnaire reported that they have requested or used more reports and test information this year than they

have in previous years. Of those interviewed, 65% claimed they have used individual item analysis, 71% have used group item analysis, and 41% have requested or used other types of test reports this year more than in previous years. It is difficult to determine how much of this was due to the program and how much was a result of changes in district testing policy.

The interview procedure with administrators did not support the professed 71% use of group item analysis data. There appeared to be considerable variability among the districts in how, or if, group item analysis was used. Most administrators indicated that if teachers used group data at all, the use was primarily out of curiosity. Part of this lack of group data use may be due to the general absence of these prepared forms during the 1967-77 school year. The questionnaire responses did reflect some desire to request these forms during the 1977-78 school year.

The introduction of administrators to group diagnostic information and the ease with which it can be interpreted may stimulate future interest among classroom teachers in its application. The interview sessions did reveal some growing interest in the potential of using group data for instructional grouping purposes. However, this came more from administrators who also saw the use of group data as a possible way of quickly identifying gross skill deficiencies. Some written comments related to the use of group

data in identifying general areas of skill deficiencies and matching test results to curriculum objectives (scope and sequence).

A major objective of the staff development program was to encourage teachers to make more effective use of distributional data. There was no evidence that teachers, as a result of the program, were using their instructional group test score distributions in identifying and providing instructional assistance to children in the asymmetrical segments of the distributions. However, the cognitive skill area of distribution characteristics (see Table 3) showed the greatest amount of change from pre- to post-assessment. One possibility for the demonstrated lack of use may be the general unavailability of these types of reports. There was considerable evidence, particularly in one district, of grouping children for instructional purposes based on individual pupil item analysis information. In another district, systematic procedures with prerequisite training were established for communicating test results to parents and children based on individual item analysis data. Consequently, use was made of the more diagnostic type of data when those data were readily available.

The final version of the staff development program was presented in one school district during the spring of 1977. This educational community was composed of 18

administrators, classroom teachers, and guidance counselors. Eighty-one percent of these participants responded that the program helped them to be more realistic in their use of test results. Seventy-five percent stated the program clarified well enough the basic statistical concepts used in testing, and 81% claimed it gave them a better understanding of the basic statistics involved. The majority also felt the program needed follow-up workshop type activity. Counselors viewed the program as desirable training for their staffs. Teachers, though generally satisfied, wanted more time with the material.

The process of developing and evaluating the staff development program in tests and measurement have implications for current and future training of educators. These implications and other considerations are further discussed in Chapter V.

C H A P T E R V

CONCLUSIONS AND RECOMMENDATIONS

The activity of testing and measurement is being questioned today more than ever before. Requests are being made from both professional educators and the lay public for the elimination of testing, particularly norm-referenced tests. Some of these requests are based on evidence of the misuse of tests or test information, rather than any intrinsic fault of the actual instruments or techniques used. Testing problems seem to emerge when the results are applied or, more frequently, incorrectly applied to making decisions about individuals or groups.

As indicators of human behavior, test results are neither good nor bad. Their usefulness or merit depends on how they are obtained, processed, interpreted and applied in the context of other information. It was the premise of this study, and supported by a review of the literature, that educators are lacking in the basic skills necessary to perform two of these basic functions: interpreting test results and applying them to make decisions about the educational environment. Dyer (1973) points to this problem and offers a partial explanation when he states:

Dyer's First Law of Information Dilution, which states that, as knowledge expands while the population of potential users of knowledge also expands, the probability approaches unity that everybody is ignorant of what anyone else knows. In other words, the great majority of test users simply does not have the time to look up or catch up or keep up with the enormous number of tests and the mountainous literature that the testmakers continue to pile up. (p. 91)

The purpose of this study was to develop a staff development program for bringing the test user closer to the testing technology. The program, in addition to upgrading the test and measurement skills of educators, was designed as an on-site, in-service training program. The content of the staff development program was designed to improve basic skills in tests and measurement necessary for effective application of most any test information. However, part of the program was also designed to be flexible enough to deal with the application of test information which would be unique to each district.

Summary of the Findings

The major strengths of the staff development program were in its presentation and interpretation of statistics and in the presentation of different types of testing activities and results. There was also considerable positive comment from the participants regarding the scope and sequence of the materials presented. A basic program objective was to prepare materials that were easy to understand

in explaining the basic information necessary for effective test interpretation. The majority of the responses, both quantitative and qualitative, supported the achievement of this objective.

A major focus of the staff development program was in its stress on instructional uses of group data. As indicated in Chapter IV, the program's emphasis on the instructional use of group data did not appear to be realized in the operational educational environments in this study. There was evidence that teachers used individual pupil item analysis data for instructional grouping. It is possible that group data was not used due to the general lack of familiarity with this type of data and the need for further training in its application. Another reason may be the more threatening nature of group data in that group skill deficiencies, particularly in the case of spring testing, can be identified by school and teacher.

The staff development program in tests and measurement made a significant contribution to the knowledge base of most of the educators who participated in its implementation. The program's impact on cognitive skill improvement was observed through pre- and post-training assessment as well as through interview responses. Pre- and post-assessment improvement was noted in the areas of knowledge and use of test data, and in items measuring reliability and validity. The skill area of greatest change, as measured

by the pre- and post-assessment instrument, was found to be in items dealing with test score distribution characteristics. Though the staff development program did not appear to stimulate greater use of group data, the evaluation information indicated the program had some impact on the participants' understanding of some basic principles underlying the use of group data.

The inclusion in the program of material designed to present some weaknesses and common misuses of test data did not appear to stimulate negative attitudes about testing. Seventy-eight percent of the participants responding to the questionnaire reported that the program helped them to make more realistic use of test data. When questioned about the term "realistic" during the interview process, the comments related to their being introduced to weaknesses and common misuses of test data. Consequently, the introduction of this information did not appear to adversely affect attitudes toward testing. The attitudes about tests and measurement of the educators in this project were relatively positive prior to the implementation of the program, and remained at about the same level through training. This information is consistent with the review of the literature (Hastings, et al., 1961; Brim, et al., 1964; Goslin, 1967; Short and Szabo, 1974; Cormany, 1974; Stuck and Wyne, 1977; Bhaerman, 1977) which also suggests that the attitudes of many educators toward tests and testing information

is relatively positive.

The program appeared to have its greatest impact on administrators. Perhaps some of this was due to the current significance placed on test information and the need for administrators to be knowledgeable as a result of a mounting community interest. Administrators are also becoming more aware of what is available through the literature they receive and the conferences they attend. Several expressed the desire to upgrade their own and their teachers' skills with what they knew was available. Whatever the reason, the administrator participants were very responsive to the program and provided some of the most constructive feedback. For example, administrators offered excellent advice regarding content for follow-up training activities.

The usefulness of the staff development program for classroom teachers appears to be based primarily on the amount of time the school district is willing to contribute to follow-up in-service activity. In general, the program presents an overview of tests and measurement and needs more follow-up activity than was provided in the course of this study. While the administrator needs the understanding an overview can provide, the classroom teacher is the one who has to apply the test information for instructional management. The classroom teacher needs more direct and relevant training and would benefit from any extension of

the application section of the program.

Educational specialists who participated in the project claimed the program served their needs in making them more constructively critical of the test information they use. Typical comments indicated that they felt they wanted to use test information as a means to an end rather than as an end in itself. These specialists generally included people in the reading or counseling area and most had some training in tests and measurement. Most viewed the program as a "refresher course" with additional insights into measurement skills, varieties of instruments, and techniques for displaying and interpreting data.

The process evaluation approach used in the staff development program made the recipients of the study serve as "participant evaluators." This dual role of both learner and evaluator seemed to enhance the general level of interest. The participants were encouraged to find fault with the program and, if possible, recommend ways that it could be improved. Most participants appeared to appreciate being a part of the activity rather than just a member of another in-service training class. Though this may be considered a delimitation of the study, which will be discussed in the following section, it could also be used as a standard approach for program implementation. The staff development program was designed to be a general training program. However, unique situations in every

school district plus the development of new tests and testing methodologies will require that the program be continually reviewed and modified. Consequently, it may be desirable to consider the "participant evaluator" approach as part of the treatment in any further development or modification of the program.

Delimitations of the Study

This study describes the formative processes in developing and evaluating an in-service staff development program. The study is not a research project involving classical experimental design. It is more closely related to what Campbell and Stanley (1963) refer to as a quasi-experimental design in that "full" experimental control is lacking. Consequently, there are threats to both internal validity -- did the program make a difference, and external validity -- could program effects be generalized to other educational environments. Several of the extraneous variables referred to by Campbell and Stanley (1963) are not applicable in this study. For example, experimental mortality is not applicable since there was no control group employed. There was no opportunity for multiple-treatment interference to occur since multiple treatments were not applied to the same respondents. The major threats to internal validity were history and maturation since there were

no controls for the activities, exclusive of the program, that took place between testing, and the length of program implementation was intentionally varied. The testing process may also have been a threat to internal validity since the participants took the same instrument both before and after training, and there was not a control group to check for testing effects. A major threat to external validity may have been possible reactive effects of using the "participant evaluator" approach to program development and evaluation.

The use of "participant evaluators" was mentioned earlier as an advantage due to the level of interest this approach appeared to stimulate. However, this also introduced an element of bias by placing the program participants into a role they may not perform outside of the context of this study. However, it may be desirable in any future use of the staff development program to encourage the participants to assume an evaluator role. The continued value of the program will be measured by how well it meets the changing needs of test users, and by how well it reflects changing testing methodology.

The involvement of the author of the staff development program in its development, implementation, and evaluation also introduced an element of bias. The pre- and post-assessment instruments of cognitive and affective skills and the questionnaires were used in an attempt to

gather objective data to offset the personal involvement in the interview process.

A problem related to generalizability was the regional nature of the sample used in field testing the staff development program. The groups participating in this program represented a small number of northeast educational communities and may not reflect the attitudes or knowledge of teachers and administrators from other parts of the country or large urban school systems. However, there was considerable similarity in the cognitive and attitudinal responses obtained from this study and those reflected in the literature (Hastings, et al., 1961 and Hotvedt, 1974).

A serious delimitation of the study involved the amount of training time available. Many school systems have strict limits on the amount of time that can be devoted to in-service training. This limitation required the development of a flexible training package which could be easily adapted to fit a variety of time constraints. However, there was still not enough time provided for follow-up activities with individual or small group participation, e.g., greater emphasis in interpreting and applying local data in meeting specific instructional needs and in reporting these data to parents. This lack of time appeared to be a consistent problem raised throughout the project. There needs to be a commitment on the part of the

school system for more time for intensive on-going training activities.

Training Considerations and Recommendations

Considerations

Evidence gained from implementing and evaluating the in-service staff development program on tests and measurement indicated that the participating educators benefited from the experience. The program had no adverse affects on attitudes about tests or testing activities. Cognitive skills related to tests and measurement were significantly increased in most of the school districts involved, and the increase in skills was more noticeable as the program was refined through the formative evaluation process. The majority of the participants responded favorably to the program's content, format, and method of presentation. Both teachers and administrators reported more realistic and more effective use of test data as a result of participating in the staff development program. One indicator of such effectiveness was the evidence that at least two districts were using data more effectively for instructional grouping and reporting to parents. What is not known, and is beyond the scope of this study, is what impact the program will have on long range instructional practices. This study has involved a developmental activity. The basic product, the

in-service staff development program, has been developed and should serve as the basis for further research on its effectiveness in a larger educational environment.

Recommendations

The following recommendations are made as a result of the implementation and evaluation of the staff development program.

1. More emphasis needs to be placed in the program on the necessity of first defining the purposes of testing. Educators need to realize that different types of tests are designed to satisfy different types of objectives. If the objectives of the testing program are specified first, then the proper test or information gathering procedure can be selected and applied.

2. The part of the staff development program that deals with the instructional application of group data should be amplified and made more specific to the unique needs of the educational community being served. The use of group data from norm-referenced standardized achievement tests for classroom instructional management appears to be less frequent than is merited by the potential of these data. A practicum provided to educators using their own group data may increase awareness and foster the use of group results.

3. The staff development program should be implemented on an introductory basis for all staff members. The initial presentation of the program should not exceed four hours and should be introductory to follow-up in-service training for teachers. Follow-up activities should incorporate the test materials and data used in the particular educational community receiving the training. Emphasis should be placed on providing simulated or actual experiences using local data for instructional decision-making at the classroom level and interpreting test information to parents.

4. Implementation of the staff development program should coincide with the local testing program. If the training coincides with the local testing program there may be more motivation to learn due to the immediacy of interpreting and applying data.

5. Copies of all the materials presented visually should be provided for all participants in the staff development program. The participants should be instructed that these materials are for their future reference and note-taking purposes. If possible, these materials should be provided a few weeks prior to the implementation of the program. Prior review was expressed by some participants as desirable for the development of questions which could

stimulate more discussion.

Further Development and Research

This study has concentrated on the development of a training program for the improvement of educators' skills in tests and measurement. The resulting program has been refined through on-site field testing. The evidence indicates that the program has been successful in improving the tests and measurement skills of educators. However, there is still a need for further development and research.

The actual materials used, the transparencies, could be made more attractive and appealing through the addition of graphic arts. The use of graphics could be examined as a variable in improving participant response to the program.

Further developmental activities could include the creation of simulation exercises to be used in the application section. Such exercises could include role playing activities involving teacher-parent interaction with test results and the application of individual and group data to instructional problems.

An additional developmental activity could involve training local school personnel in implementing the program in their own school districts. The implementation of the staff development program was accomplished by the author of the program. In order to respond to the question

of generalizability, the program should be effectively administered by others in a variety of educational environments. Furthermore, some of the program may serve as an effective vehicle for informing the lay public -- such as boards of education -- about more effective uses of test information. Further research should explore the variety of audiences for which this program, or parts of it, may serve different segments of the educational community.

Changes in the testing industry will require changes in the staff development program. Provisions should be made for the inclusion and field testing of new material.

The basic research question that needs to be answered is what effect this program has on the long-range improvement of instructional practices and the associated effects it may have on improving the basic skills of children. Fundamental to this question is the issue of how teachers use data. Hotvedt (1974) addressed the issue of test use through a case study approach in a school district. The results indicated that test use was a function of many variables, most of which seemed to be related to the availability, type, and timing of data. These also seemed to be important variables in this study, but not as important as the overall issue of training. Immediate usage of data in the districts studied appeared to be related to learning more about what data were available and how they could be more effectively used. A major problem in determining

specifically how teachers use data is in obtaining accurate information. Hotvedt (1974) found this a major difficulty in his case study and proposed a possible solution through observation. It may be desirable to spend time in a school district, observing and taking field notes on the use of test data prior to the implementation of the staff development program. In this way, the program can be adapted to emphasize the deficiencies noted in test data use. An important question for future research involves determining what types of test data produce the most significant instructional improvement. The answer may be different for each school district, school or specific instructional environment. However, once determined, issues of training, testing time, and availability of data would be greatly simplified.

A point made early in the staff development program is that "testing" and "evaluation" are not synonymous terms. Evaluation is a process leading to decision-making, and as such, is a highly subjective activity relying on a variety of sources of information. Testing can serve as one of those sources of information in the process of decision-making, but cannot and should not be the only source. The staff development program described in this study is designed to place the activity of testing and its results into a proper educational perspective. When testing is understood and its results are used in the context of other information, it should play a significant role in helping teachers teach and learners learn.

BIBLIOGRAPHY

- Airasian, P. W. and Madaus, G. F. A Study of the Sensitivity of School and Program Effectiveness Measures. Report submitted to the Carnegie Corporation of New York. Boston, Mass.: Boston College, 1976.
- Amos, J. R.; Brown, F. L.; and Mink, O. G. Statistical Concepts: A Basic Program. New York, N. Y.: Harper & Row Publishers, 1965.
- Asher, J. W. Educational Research and Evaluation Methods. Boston, Mass.: Little, Brown and Company, 1976.
- Backman, M. E. "Teacher-Made Tests: The Test Director's Role," in A. Ward, M. E. Backman, B. W. Hall and J. L. Mazur (Eds.), Guide for School Testing Programs (NCME). East Lansing, Michigan: Michigan State University, 1976.
- Bhaerman, R. D. "What Do Teachers Think About Testing?" American Educator, 1977, 1, 1, 10-14.
- Bloom, B. S. "Toward a Theory of Testing Which Includes Measurement - Evaluation - Assessment," in M. C. Wittrock and D. E. Wiley (Eds.), The Evaluation of Instruction: Issues and Problems. New York, N. Y.: Holt, Rinehart and Winston, Inc., 1970.
- Bloom, B. S.; Hastings, J. T.; and Madaus, G. F. Handbook on Formative and Summative Evaluation of Student Learning. New York, N. Y.: McGraw-Hill, 1971.
- Brady, E. "To Test or Not To Test." American Educator, 1977, 1, 1, 3-9.
- Brim, O. G.; Goslin, D. A.; Glass, D. C.; and Goldberg, I. The Use of Standardized Ability Tests in American Secondary Schools and Their Impact on Students, Teachers, and Administrators. (Russell Sage Foundation Technical Report 3). New York, N. Y.: Russell Sage Foundation, 1964.
- Bruning, J. L. and Kintz, B. L. Computational Handbook of Statistics. Glenview, Ill.: Scott, Foresman & Co., 1968.

- Campbell, D. T. and Stanley, J. C. Experimental and Quasi-Experimental Designs for Research. Chicago, Ill.: Rand McNally College Publishing Co., 1963.
- Cooley, W. W. Methods of Evaluating School Innovations. Pittsburgh, Pa.: Learning Research and Development Center, 1971. ✓
- Cormany, R. B. "Faculty Attitudes Toward Standardized Testing." Measurement and Evaluation in Guidance, 1974, 7, 188-194.
- Cronbach, L. J. Essentials of Psychological Testing. New York, N. Y.: Harper & Row, 1970.
- Dubois, P. H. "A Test-Dominated Society: China, 1115 B.C. - 1905 A.D." Proceedings of the 1964 Invitational Conference on Testing Problems. Princeton, N. J.: Educational Testing Service, 1964, pp. 3-11.
- Dyer, H. S. "The Role of Evaluation in School Systems." A paper presented at the New Jersey Association of School Administrators, Atlantic City, New Jersey, 1971.
- Dyer, H. S. "The Role of Evaluation in Curriculum Innovation," in A. Dragositz (Ed.), Proceedings of the Association for Supervision and Curriculum Development Pre-Conference Seminar. Princeton, N. J.: Educational Testing Service, 1969.
- Ebel, R. L. "The Social Consequences of Educational Testing," in D. A. Payne and R. F. McMorris (Eds.), Educational and Psychological Measurement: Contributions to Theory and Practice. Waltham, Mass.: Blaisdell Publishing Co., 1967.
- Fleming, M. "Standardized Tests Revisited." School Counselor, 1971, 19, 2, 71-72.
- Gellman, E. S. Statistics for Teachers. New York, N. Y.: Harper & Row Publishers, 1973.
- Glaser, R. "Instructional Technology and the Measurement of Learning Outcomes: Some Questions." American Psychologist, 1963, 18, 519-521.
- Goslin, D. A. Teachers and Testing. New York, N. Y.: Russell Sage Foundation, 1967.

- Grant, J. W. Personal communication, August 5, 1976.
- Hambleton, R. K. "Principles of Educational and Psychological Testing." (Course materials) Amherst, Mass.: University of Massachusetts, Amherst, 1974.
- Hambleton, R. K. and Gorth, W. P. Criterion-Referenced Testing: Issues and Applications. Amherst, Mass.: University of Massachusetts, Tr. No. 13, 1971.
- Harnischfeger, A. and Wiley, D. E. "Achievement Score Decline: Do we Need to Worry?" Monograph of the ML-Group Studies in Education, Cemrel, Inc., 1975.
- Harrington, C. Measurement Concepts: A Basic Program. New York, N. Y.: Vantage Press, 1968.
- Harris, C. W. (Ed.) Problems of Measuring Change. Madison, Wisconsin: The University of Wisconsin Press, 1967.
- Hastings, J. T.; Runkel, P. J.; and Damrin, D. E. Changes in Schools Which Do and Do Not Send Staff Members to Training Institutes in Counseling. (University of Illinois Bureau of Educational Research CRP-939). Urbana, Ill.: University of Illinois, 1961.
- Hodges, J. L.; Krech, D.; and Crutchfield, R. S. STATLAB: An Empirical Introduction to Statistics. New York, N. Y.: McGraw-Hill Book Co., 1975.
- Hotvedt, M. O. "A Case Study of Standardized Test Use in the Public Schools." Unpublished Ph.D. dissertation, University of Illinois, 1974.
- Huck, S. W.; Cormier, W. H.; and Bounds, W. G. Reading Statistics and Research. New York, N. Y.: Harper & Row Publishers, 1974.
- Jackson, P. W. Life in the Classrooms. New York, N. Y.: Holt, Rinehart and Winston, 1968. ✓
- Kirby, J. H.; Culp, W. H.; and Kirby, J. MUST: Manuals for Users of Standardized Tests. Bensenville, Ill.: Scholastic Testing Service, Inc., 1973.
- Klitgaard, R. E. "Going Beyond the Mean in Educational Evaluation." P-5184, The Rand Corporation, March 1974.
- Ladd, E. M. "More Than Scores From Tests." Reading Teacher, 1971, 24, 4, 305-311.

- Lindvall, C. M. and Nitko, A. J. Measuring Pupil Achievement and Aptitude. New York, N. Y.: Harcourt, Brace and Jovanovich, 1975.
- Lohnes, P. R. "Statistical Descriptors of School Classes." American Educational Research Journal, 1972, 9, 547-556.
- Lyman, H. B. Test Scores and What They Mean. Englewood Cliffs, N. J.: Prentice-Hall, Inc., 1971.
- Mayo, S. T. Trends in the Teaching of the First Course in Measurement. Chicago, Ill.: Loyola University, 1970. (ERIC Document Reproduction Service No. ED 047 007.)
- McKenna, B. "What's Wrong With Standardized Testing?" Today's Education, 1977, 66, 2, 35-38.
- National Council on Measurement and Evaluation. Guide for School Testing Programs, 1976.
- National School Board Association. Standardized Achievement Testing. National School Board Association Report 1977-1. Washington, D. C.: National School Board Association, 1977.
- Ozenne, D. G. Toward an Evaluative Methodology for Criterion-Referenced Measures: Test Sensitivity. Los Angeles, Calif.: University of California at Los Angeles, CSE Report No. 72, 1971.
- Popham, W. J. Criterion-Referenced Measurement: An Introduction. Englewood Cliffs, N. J.: Prentice-Hall, Inc., 1975.
- Provus, M. Discrepancy Evaluation. Berkeley, Calif.: McCutchan Publishing Co., 1971.
- Randall, R. S. "Contrasting Norm-Referenced and Criterion-Referenced Measures." A paper presented at the American Educational Research Association, Chicago, Illinois, April 1972.
- Resolutions. National Education Association Publications. July 1972, p. 36.
- Sax, G. "The Use of Standardized Tests in Evaluation," in W. J. Popham (Ed.), Evaluation in Education: Current Applications. Berkeley, Calif.: McCutchan Publishing, 1974.

- Scriven, M. "The Methodology of Evaluation," in R. W. Tyler, R. M. Gagne, and M. Scriven (Eds.), Perspectives of Curriculum Evaluation. Chicago, Ill.: Rand McNally, 1967.
- Short, B. G. and Szabo, M. "Secondary School Teachers' Knowledge of and Attitudes Toward Educational Research." Journal of Experimental Education, 1974, 43, 75-78.
- Smith, C. W. "Criterion-Referenced Assessment in Contrast to Norm-Referenced Measurement." A paper presented at the International Symposium on Educational Testing, The Hague, The Netherlands, July 1973.
- Stuck, G. B. and Wyne, M. D. "Teachers' and Pupils' Opinions Relative to the Major Issues in Collecting and Using Pupil Test Data." Paper presented at the meeting of the American Educational Research Association, New York, April 1977.
- "Teacher-Made Tests." Today's Education, March-April 1977, pp. 52-53.
- TenBrink, T. D. Evaluation: A Practical Guide for Teachers. New York, N. Y.: McGraw-Hill, 1974.
- Traxler, A. E. "Fifteen Criteria of a Testing Program," in D. A. Payne and R. F. McMorris (Eds.), Educational and Psychological Measurement: Contributions to Theory and Practice. Waltham, Mass.: Blaisdell Publishing Co., 1967.
- Tuckman, B. W. Measuring Educational Outcomes: Fundamentals of Testing. New York, N. Y.: Harcourt, Brace, Jovanovich, Inc., 1975.
- Tukey, J. W. and Wilk, M. B. "Data Analysis and Statistics: An Expository Overview." Proceedings of the Fall Joint Computer Conference. 1966, 695-709.
- Warries, E. "Standard Mastery Curves and Skew Curves." A paper presented at the Annual Meeting of the American Educational Research Association, Chicago, 1974.
- Wick, J. W. Educational Measurement: Where Are We Going and How Will we Know When we Get There? Columbus, Ohio: Charles E. Merrill Publishing Co., 1973.

APPENDIX A

Pre- and Post-Assessment Instrument

Number _____

INSTRUCTIONS

The attached looks like a test, and in one way, it is. But the purpose in asking you to try your hand at it is quite different from the purpose which many tests are given.

This is not a measure of your ability or your competence.

The intent of this instrument is to find out the kinds of things which you remember and use about tests and measurement as practicing teachers. The major purpose is to determine the kinds of technical information about testing which are most salient.

Please respond by circling the letter that you feel corresponds to the best answer. Mark only one answer for each question and try to avoid dwelling too long on any one item.

Thank you for your cooperation.

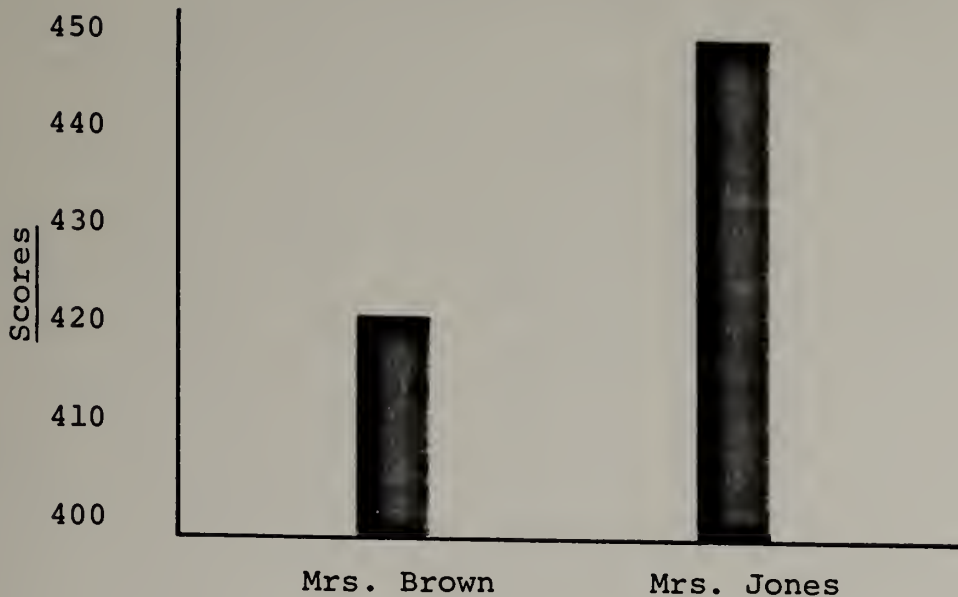
Note: The term "standardized test" used in this document will refer to norm-referenced standardized tests, such as the Iowa Test of Basic Skills or the Comprehensive Tests of Basic Skills.

1. A student scored at the 75th percentile on a standardized achievement test. This means that
 - A. 75 percent of the norm group scored lower than the student.
 - B. 75 percent of the norm group scored higher than the student.
 - C. The student answered 75 percent of the questions correctly.
 - D. The student is in the upper 25 percent of his own high school class.

2. Because no standardized test possesses perfect reliability, it is essential that the teacher regard the score which a student obtains as
 - A. Having little meaning unless it is very high or very low.
 - B. Indicating a point in the range near which the student's true score probably falls.
 - C. Indicating only that the student has either more, or less, ability than the average individual in the norming group.
 - D. Providing information about the student which can be used only by a thoroughly trained school psychologist.

3. Objective measures of performance may have a negative effect on learning.
 - A. Agree
 - B. Partially agree
 - C. No opinion
 - D. Partially disagree
 - E. Disagree

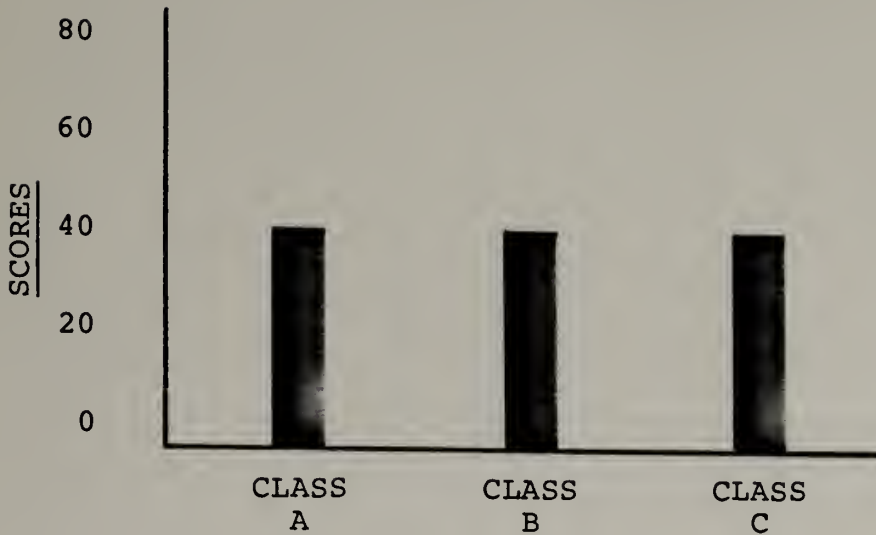
Grade 6 Average Scores on the
Comprehensive Tests of Basic Skills



4. According to the graph above which shows the spring test results on the Comprehensive Tests of Basic Skills, Mrs. Brown's class has a lower average score than Mrs. Jones' class. This difference most certainly indicates that
- A. Mrs. Brown devoted less time to individual instruction than Mrs. Jones.
 - B. Mrs. Jones devoted more time to instruction in the areas measured by the test.
 - C. Mrs. Jones' class has a wider range of scores than Mrs. Brown's class.
 - D. A wide range of ability exists among Mrs. Jones' pupils.
 - E. A difference in general academic achievement exists between Mrs. Brown's and Mrs. Jones' pupils.

5. If a test has a mean of 50 and a standard deviation of 10, approximately two-thirds of the group received scores between
- A. 40 and 50
 - B. 40 and 60
 - C. 50 and 60
 - D. 30 and 70
6. Group data from standardized achievement tests have little value in instructional planning for individual children.
- A. Agree
 - B. Partially agree
 - C. No opinion
 - D. Partially disagree
 - E. Disagree
7. Early in the school year, a teacher should receive intelligence and achievement test scores for his/her pupils.
- A. Agree
 - B. Partially agree
 - C. No opinion
 - D. Partially disagree
 - E. Disagree
8. Which one of the following statements most closely matches your opinions about the use of standardized test results?
- A. They have limited value because they cause a teacher to "categorize" students.
 - B. They provide information on which to base further study.
 - C. They can provide information for classroom instructional management.
 - D. They are too unreliable to be of value except for drawing conclusions about groups of individuals.

CLASS AVERAGES ON READING SKILLS
TEST



9. The graph above indicates that as far as the reading skills measured by this test given in the fall are concerned:
- The materials used in these classes should have approximately the same difficulty level.
 - The similarity of these classes would be important in testing the long term effects of differing instructional practices.
 - The children in the top or bottom quarter of the scoring range may differ with respect to level of reading skill mastery.
 - The range and diversity of reading skills is about the same for all classes.

10. Year after year, the mean achievement test scores for the students in School X consistently fall one year or more above the national norms. What is the most probable cause of this finding?
- A. School X is located in an upper-middle class community.
 - B. School X is staffed with expert teachers.
 - C. School X is using tests that have unreliable norms.
 - D. School X stresses the traditional, rather than a more process-oriented curriculum.
11. School Y's grade 6.0 had a mean total battery standard score on the Comprehensive Tests of Basic Skills of 504. The percentile tables in the manual indicate that this standard score was equal to a percentile rank of 74. This indicates that
- A. Grade 6 children in School Y answered 74% of the items correctly.
 - B. School Y did better than 74% of the schools in the national sample.
 - C. Children in School Y as a whole did better than 74% of other children tested in the same grade level.
 - D. The majority of children in School Y performed above the national average.
12. A valid test is one which
- A. Has good item discrimination indices.
 - B. Measures what it is supposed to measure.
 - C. Has a relatively large number of items.
 - D. Correlates well with I.Q. tests.

13. The reliability of a test is determined by
- A. The length of the test.
 - B. The correlation of the test with a retest situation.
 - C. How well it measures what it is supposed to measure.
 - D. The difficulty level of the items.
14. During the first week of school, a teacher gave the same reading comprehension test to both of her fifth grade English classes. In her morning class the students had a mean of 73 and a standard deviation of 12. In her afternoon class, the students had a mean of 73 and a standard deviation of 26. What do these results imply with regard to her planning for these two classes?
- A. The difficulty range of reading materials should be greater for the afternoon class.
 - B. The variety of reading materials should be greater for the afternoon class.
 - C. The textbook used in the afternoon class should be of simpler and easier format.
 - D. The work assignments given the afternoon class should be less extensive.
15. The idea of a yearly testing program of abilities or aptitude is a good one.
- A. Agree
 - B. Partially agree
 - C. No opinion
 - D. Partially disagree
 - E. Disagree

16. The idea of a yearly testing program of achievement in basic skill areas is a good one.

- A. Agree
- B. Partially agree
- C. No opinion
- D. Partially disagree
- E. Disagree

17. Below is a list of different kinds of tests. On the lines to the right of each, please place a check mark if you are not familiar with it or if your school gives and/or ought to give it to pupils.

	I am not familiar with this type of test	This type of test is presently being given	This type of test ought to be given	This type of test should not be given
	_____	_____	_____	_____
Intelligence tests	_____	_____	_____	_____
Academic aptitude tests	_____	_____	_____	_____
Norm-referenced standardized achievement tests	_____	_____	_____	_____
Criterion-referenced achievement tests	_____	_____	_____	_____
Interest tests	_____	_____	_____	_____

18. About how many of the other teachers in your school do you think would check the tests you did? (Make the best guess you can if you are not sure.)

- _____ Almost all
- _____ More than half
- _____ Less than half
- _____ Almost none

19. With which one of the following statements would you tend to agree?
- A. A test score is a valuable piece of information which answers many of the questions I have about my students.
 - B. A test score is a valuable piece of information useful in raising important questions in my mind about my students.
 - C. A test score is an interesting piece of technical information but possesses little or no value for the on-going activities of the classroom.
20. In talking with other teachers about students, do you discuss the results of standardized tests with them?
- Frequently
 - Sometimes
 - Rarely
 - Never
21. How much formal training have you had in college in the area of tests and measurement?
- More than three courses
 - Three courses
 - Two courses
 - One course
 - None
22. How much training have you had in tests and measurement since you left college? (Consider courses as classes or number of workshops, institutes, inservice training sessions, etc.)
- More than three courses
 - Three courses
 - Two courses
 - One course
 - None

APPENDIX B

Questionnaire Form

EVALUATION QUESTIONNAIREON THE STAFF DEVELOPMENT PROGRAM ON TESTS AND MEASUREMENTFALL 1976

This fall you participated in a staff development program on tests and measurement. This program was designed to impart basic knowledge about achievement tests and measurement to educators in an effort to present and clarify some of the current testing issues and make achievement test results more useful in the instructional environment. This questionnaire is provided in the hope that you will respond constructively about the ways you think the staff development program can be strengthened or modified to better meet the needs of program participants. Unless specified otherwise, the questions refer primarily to norm-referenced standardized achievement testing.

If you feel that some questions cannot be answered with a simple "yes" or "no", please place a check mark in the column titled "Other" and explain your reason on the reverse side of the page. If you are an administrator you may not have a personal involvement in the direct application of test data to the instructional environment of individual children. If this is the case, please respond to these types of items as you perceive the way teachers are currently using test information.

Your response to the questionnaire is strictly voluntary, anonymous, and sincerely requested. A self-addressed stamped envelope is provided for its return. Thank you for your cooperation in the training sessions and in completing this questionnaire.

I am a:

Classroom Teacher _____

Administrator _____

Specialist _____

I attended the:

First Session _____

Second Session _____

Third Session _____

- | | <u>Yes</u> | <u>No</u> | <u>Other</u> |
|---|------------|-----------|--------------|
| 1. How do you feel about achievement testing in general? More specifically: | | | |
| Do you look at previous year's results in helping you make instructional decisions about these same children this year? | — | — | — |
| Do you think test results from previous years could give you the wrong impression about your pupil's performance? | — | — | — |
| Based on what you know about your children, do you think this year's test results are accurate? | — | — | — |
| Do you think achievement testing can help you identify skill deficiencies in your pupils that may not have been apparent in classroom activities? | — | — | — |
| Do you think you are adequately trained for the level of test interpretation and use that is necessary in your particular situation? | — | — | — |
| If not, in what specific skills would you like to have further training? _____ | | | |
| _____ | | | |
| _____ | | | |
| 2. If your school district has a mandated testing program, what do you think about the time involved in testing? | | | |
| The amount of time spent in testing is about right. _____ | | | |
| There is too much time spent in testing. _____ | | | |
| There is not enough time spent in testing. _____ | | | |

3. If your school district has a mandated testing program, what kind of test information are you now getting:

For individual pupils _____

For your classroom _____

For your school _____

For your district _____

4. If there were no cost restrictions on testing, and you could get anything you wanted in the way of test results, what would be the most desirable information you could receive:

For individual pupils _____

For your classroom _____

For your school _____

For your district _____

- | | <u>Yes</u> | <u>No</u> | <u>Other</u> |
|--|------------|-----------|--------------|
| 5. Do you get involved in the process of administering norm-referenced aptitude or achievement tests that are mandated in your district? | --- | --- | --- |
| If yes, in what way do you get involved? _____ | | | |
| _____ | | | |
| _____ | | | |
| 6. Based on what you currently know about achievement testing, do you think it is a worthwhile activity for gaining useful information about children? | --- | --- | --- |
| If yes, what current information have you found most useful? _____ | | | |
| _____ | | | |
| _____ | | | |
| If no, why not? _____ | | | |
| _____ | | | |
| _____ | | | |
| 7. Do you use the test publishers' manuals in helping you analyze or interpret test data? | --- | --- | --- |
| If yes, which one(s) are most useful? _____ | | | |
| _____ | | | |
| _____ | | | |
| Have you used these kinds of manuals more this year than in previous years? _____ | | | |

Yes No Other

8. Have you used a frequency distribution or a distribution of scores as an aid in determining instructional grouping or use?

Prior to this school year

This school year

9. Have you requested or used other types of reports from norm-referenced achievement tests more this year than in previous years?

If yes, which ones and why? _____

10. Did the staff development program on tests and measurement that you attended this fall:

Give you sufficient information for your needs?

Give you more information than you needed?

Give you less information than you needed?

If yes, what should be added to make it more useful to you? _____

	<u>Yes</u>	<u>No</u>	<u>Other</u>
11. Do you feel the staff development program:			
Presented information too quickly -- needed more time for discussion?	---	---	---
Presented information too quickly -- needed more elaboration and expansion over a greater period of time?	---	---	---
Presented information too slowly -- could have covered it in less time?	---	---	---
Needed more workshop sessions to discuss actual use of individual pupil or group data?	---	---	---
Made you critical of testing in general?	---	---	---
Helped you to be more realistic in your use of test results?	---	---	---
Did not clarify well enough the basic statistical concepts used in testing?	---	---	---
Gave you a better understanding of the basic statistics used in testing?	---	---	---

Please explain any of your responses that could contribute toward improving the quality of the program or add any comments that may not be covered above.

APPENDIX C

Interview Form

INTERVIEW FORM

Yes No

1. What kind of courses have you had in tests and measurement and statistics either in or out of college?

2. How do you feel about achievement testing in general?

Do you use previous year's test results in helping you make instructional decisions about this year's pupils? _____

Do you think test results from previous years could give you the wrong impressions about your pupil's performance? _____

Based on what you know about your children, do you think this year's test results are accurate? _____

Do you think achievement testing can help you identify skill deficiencies in your pupils that may not have been apparent in classroom activities? _____

Do you feel the level of statistics covered in the staff development program was too complex or involved for your needs? _____

Do you think there is too much testing in your school? _____

Do you think there is not enough of the right kind of testing in your school? _____

If yes, what would you envision as the right kind of testing? _____

Yes No

3. What is your position regarding the relationship between norm-referenced tests and criterion-referenced tests?

Norm-referenced tests are sufficient on their own.

Criterion-referenced tests are more useful and should replace norm-referenced tests.

Norm-referenced tests are useful and should be supplemented with criterion-referenced testing information.

4. How much time do you spend in testing your pupils each year (not including the time spent in formal town-wide or mandated testing programs)?

Approximately _____ minutes.

5. How much time are your children involved each year in formal town-wide or mandated testing programs?

Approximately _____ minutes.

6. Have you looked at pupils' answers to particular items on the test more this year than in previous years?

If yes, why and how? _____

Yes No

7. Have you looked at group item response information from the test more this year than in previous years?

8. Have you requested or used other types of reports from norm-referenced achievement tests more this year than in previous years?

If yes, which ones and why? _____

9. How do you use reports on individual pupils in helping you diagnose skill deficiencies?

10. Do you discuss these reports with children and parents?

Children

Parents

APPENDIX D

Pre- and Post-Assessment Results of the
Non-Cognitive Skill Test Items

PERCENTAGE OF PRE AND POST-ASSESSMENT RESPONSES ON THE
ATTITUDINAL ITEMS
(n = 112)

<u>Items</u>	<u>Agree</u>	<u>Part- ially Agree</u>	<u>No Opinion</u>	<u>Part- ially Dis- agree</u>	<u>Dis- agree</u>
3. Objective measures may have a negative effect on learning.					
Pre	11	42	11	11	25
Post	14	34	5	10	36
6. Group data from standardized achievement tests have little value in instructional planning for individual children.					
Pre	34	22	1	13	29
Post	37	22	0	13	27
7. Early in the school year, a teacher should receive intelligence and achievement test scores for his/her pupils.					
Pre	43	23	2	13	19
Post	46	24	2	13	14
15. The idea of a yearly testing program of abilities or aptitude is a good one.					
Pre	38	34	2	7	16
Post	46	27	4	11	12
16. The idea of a yearly testing program of achievement in basic skill areas is a good one.					
Pre	64	22	1	7	4
Post	61	26	1	5	4

PERCENTAGE OF PRE AND POST-ASSESSMENT RESPONSES ON THE
 ATTITUDINAL ITEMS
 (Con't)

19. With which one of the following statements would you tend to agree?

	<u>Pre</u>	<u>Post</u>
A. A test score is a valuable piece of information which answers many of the questions I have about my students.	12	4
B. A test score is a valuable piece of information useful in raising important questions in my mind about my students.	79	88
C. A test score is an interesting piece of technical information but possesses little or no value for the on-going activities of the classroom.	5	5

PERCENTAGE OF RESPONSES REPORTED ON THE USE OF TESTS
(n = 112)

	I am not familiar with this type of test	This type of test is presently being given	This type of test ought to be given	This type of test should not be given
Intelligence tests	3	65	24	6
Academic Aptitude tests	4	71	21	4
Norm-referenced Standardized Achievement tests	10	78	11	3
Criterion-referenced Achievement tests	20	47	29	1
Interest tests	16	33	39	4

When asked how many of the other teachers in their district would respond in the same manner, the following results were recorded:

Almost all	23%
More than half	46%
Less than half	23%
Almost none	4%

When asked if teachers discussed the results of students' standardized test scores with other teachers, the following responses were recorded:

Frequently	15%
Sometimes	50%
Rarely	28%
Never	3%

PERCENTAGE OF RESPONSES ON THE BACKGROUND INFORMATION
ITEMS
(n = 112)

Item
No.

21. How much formal training have you had in college
in the area of tests and measurement?

More than three courses	4%
Three courses	14%
Two courses	23%
One course	46%
None	11%

22. How much training have you had in tests and
measurement since you left college? (Consider
courses as classes, number of workshops,
institutes, in-service training sessions, etc.)

More than three courses	3%
Three courses	5%
Two courses	9%
One course	27%
None	55%

APPENDIX E

Quantitative Results of the Interview

RESPONSES TO INTERVIEW FORM

	PERCENT	
	<u>Yes</u>	<u>No</u>
Do you use previous year's test results in helping you make instructional decisions about this year's pupils?	82	18
Do you think test results from previous years could give you the wrong impression about your pupil's performance?	83	17
Based on what you know about your children, do you think this year's test results are accurate?	88	12
Do you think achievement testing can help you identify skill deficiencies in your pupils that may not have been apparent in classroom activities?	84	16
Do you feel the level of statistics covered in the staff development program was too complex or involved for your needs?	33*	67
Do you think there is too much testing in your school?	6	94
Do you think there is not enough of the right kind of testing in your school?	48	52
Norm-referenced tests are sufficient on their own.	10	90
Criterion-referenced tests are more useful and should replace norm-referenced tests.	57	43
Norm-referenced tests are useful and should be supplemented with criterion-referenced testing information.	93	7

* Over half of the 33% came from the first district receiving the program. Several changes were made following that presentation.

	PERCENT	
	<u>Yes</u>	<u>No</u>
Have you looked at pupils' answers to particular items on the test more this year than in previous years?	65	35
Have you looked at group item response information on the test more this year than in previous years?	71	29
Have you requested or used other types of reports from norm-referenced achievement tests more this year than in previous years?	41	59
Do you discuss these results with:		
Children	70	30
Parents	98	2

APPENDIX F

Quantitative Results of the Questionnaire

RESPONSES TO THE QUESTIONNAIRE
FORM

	<u>Number Responding</u>	Percent of Responses		
		<u>Yes</u>	<u>No</u>	<u>Other</u>
Classroom teacher	21			
Administrator	35			
Specialist	19			
Unclassified	2			
Do you look at previous year's test results in helping you make instructional decisions about these same children this year?		78	21	1
Do you think test results from previous years could give you the wrong impression about your pupil's performance?		71	17	12
Based on what you know about your children, do you think this year's test results are accurate?		70	9	21
Do you think achievement testing can help you identify skill deficiencies in your pupils that may not have been apparent in classroom activities?		76	16	8
Do you think you are adequately trained for the level of test interpretation and use that is necessary in your particular situation?		79	18	3
The amount of time spent in testing is about right.		71		
There is too much time spent in testing.			4	
There is not enough time spent in testing.				12
Do you get involved in the process of administering norm-referenced aptitude or achievement tests that are mandated in your district?		46	45	9

	Percent of Responses		
	<u>Yes</u>	<u>No</u>	<u>Other</u>
Based on what you currently know about achievement testing, do you think it is a worthwhile activity for gaining useful information about children?	78	13	9
Do you use the test publishers' manuals in helping you analyze or interpret test data?	49	37	14
Have you used these kinds of manuals more this year than in previous years?	17	55	
Have you used a frequency distribution or a distribution of scores as an aid in determining instructional grouping or use?			
Prior to this school year	29	61*	
This school year	32	43*	
Have you requested or used other types of reports from norm-referenced achievement tests more this year than in previous years?	16	68	
Did the staff development program on tests and measurement that you attended this fall:			
Give you sufficient information for your needs?	66	16	10
Give you more information than you needed?	16	38	
Give you less information than you needed?	13	34	

* Two school districts do not get these reports.

	Percentage of Responses		
	<u>Yes</u>	<u>No</u>	<u>No Response</u>
Do you feel the staff development program:			
Presented information too quickly -- needed more time for discussion.	43	41	16
Presented information too quickly -- needed more elaboration and expansion over a greater period of time.	46	43	11
Presented information too slowly -- could have covered it in less time.	5	79	16
Needed more workshop sessions to discuss actual use of individual pupil or group data.	58	33	9
Made you critical of testing in general.	34	53	13
Helped you to be more realistic in your use of test results.	78	14	8
Did not clarify well enough the basic statistical concepts used in testing.	25	63	12
Gave you a better understanding of the basic statistics used in testing.	82	12	6

APPENDIX G

Changes in Program Content as a Result of
Participant Feedback

PART I - METHODS

<u>Change Made Prior to District</u>	<u>Transparency Number</u>	<u>Change Made</u>
A	3	Definitions made more concise.
A	11	Simplified criterion-referenced measurement definitions.
A	-	Eliminated derivation of I.Q.
B	-	Eliminated proper uses of norm-referenced tests as a transparency. Included as part of narrative.
B	-	Moved discussion of I.Q. to Part III.
C	3	Revised definitions.
C	5	Expanded definition of Testing.
C	8	Simplified Norm-referenced measurement definitions.
C	13	Amplified differences between Norm-referenced and criterion-referenced measurement.
C	17	Added examples of norm-referenced test results.
D	6	Assessment process definitions made more complete and concise.
D	7	Added Formative and Summative evaluation definitions.
D	18	Clarified definitions.
D	-	Moved examples of criterion-referenced tests from Part III to Part I.
E	28	Added teacher-made tests.

PART II - TECHNIQUES


<u>Change Made Prior to District</u>	<u>Transparency Number</u>	<u>Change Made</u>
A	-	Dropped bar graph of standard deviation.
C	7	Standard deviation expanded to three transparencies.
C	10	Normal curve changed and expanded to include grade equivalent scale example.
C	-	Moved "Types of Scores" to Part III.
C	-	Moved Correlation from Part III to Part II.
C	-	Amplified example of range of Correlation Coefficient.
C	-	Moved Reliability and Validity from Part III to Part II.
C	-	Redefined Validity.

PART III - APPLICATION

<u>Change Made Prior to District</u>	<u>Transparency Number</u>	<u>Change Made</u>
A	-	Changed examples of graphic displays.
C	23	Added example of frequency distribution with similar averages (means) but different distribution characteristics.
D	-	Moved Standard Score from Part II to Part III.
E	4	Clarified statements dealing with using grade equivalent scores.
E	15	Added further interpretation of Standard Error.

APPENDIX H

Final Version of the Staff Development Program



PATHS
TO
INSTRUCTIONAL
DECISIONS

USE OF TEST DATA FOR THE
IMPROVEMENT OF INSTRUCTION

© 1977 CHARLES J. CLOCK, JR.

INTRODUCTION

This in-service training program has been developed on the premise that most educational administrators and teachers do not have the required skills to understand and effectively use data derived from testing activities in diagnosing and improving instructional practices. This premise is supported by recent studies which have found teachers have not received training, in college or since, to allow them to understand and effectively use the results of norm-referenced standardized tests. The use of norm-referenced test data precludes a basic understanding of measurement terminology and statistics. Many measurement concepts that are assumed to be known by people in education are either not known or frequently misunderstood. Statistical terms used in reporting test results are not completely understood by many classroom teachers and administrators. Many teachers who have little or no knowledge of test construction techniques are constructing their own tests and making decisions about the instructional future of children based on their results. Consequently, the chances are good that incorrect decisions can be made on the basis of inaccurate, misunderstood, or inappropriately applied information. This training program is meant to impart some of the basic skills necessary for more optimal use of test data. Its emphasis is on a better understanding and use of norm-referenced standardized test data.

- METHODS - DESCRIBING THE INSTRUCTIONAL ENVIRONMENT

- TECHNIQUES - HIGHLIGHTING POTENTIAL PROBLEM AREAS

- APPLICATION - IMPROVING INSTRUCTION THROUGH UNDERSTANDING
AND APPLYING DATA

METHODS

DESCRIBING THE INSTRUCTIONAL ENVIRONMENT

SLIDE 3

Many educators use the terms of measurement, testing, assessment and evaluation interchangeably. However, these terms have different meanings and involve different processes. The major point to be stressed with this transparency is the difference between testing and evaluation - testing being an activity, and evaluation a process. Testing involves the application of an instrument or instruments to obtain information about an individual or a group. Evaluation is the application of judgement to the results of an assessment process. Evaluation involves examining information which is relevant to a particular problem in order to make effective decisions about potential solutions.

MEASUREMENT

THE ACT OF OBSERVING A BEHAVIOR OR CHARACTERISTIC OF AN INDIVIDUAL OR GROUP AND RECORDING THE INFORMATION USUALLY, BUT NOT ALWAYS, IN NUMERICAL TERMS.

TESTING

A TYPE OF MEASUREMENT ACTIVITY IN WHICH SPECIFIC INSTRUMENTS ARE USED TO DETERMINE INDIVIDUAL OR GROUP PERFORMANCE CHARACTERISTICS.

ASSESSMENT

A COMPREHENSIVE PROCESS INVOLVING THE SPECIFICATION OF A PROBLEM, THE IDENTIFICATION OF VARIABLES THAT CAN EFFECT THE PROBLEM, AND THE USE OF MEASUREMENT TECHNIQUES TO GATHER INFORMATION FOR THE EVALUATION PROCESS.

EVALUATION

A JUDGMENTAL PROCESS APPLIED TO THE RESULTS OF ASSESSMENT FOR INSTRUCTIONAL DECISION-MAKING.

SLIDE 4

Measurement is defined in a global sense and is related to the observable.

Measurement can be more effective if it involves systematic observation. An

important step in measuring anything requires a specific statement of what

and how something is to be measured. Measurement alone is not evaluation.

It can lead to evaluation because it plays an important role in the overall

assessment process.

MEASUREMENT

ANYTHING THAT CAN BE OBSERVED CAN BE MEASURED.

SYSTEMATIC OBSERVATION:

- DEFINE THE PROBLEM
- SPECIFY THE OBSERVATIONAL CRITERIA
- OBSERVE
- RECORD RESULTS

SLIDE 5

Testing is an activity which falls under the category of measurement. Testing usually involves the use of paper and pencil tests, though it can include certain types of observational techniques. Norm-referenced tests, such as the ones shown in the examples, are currently the most widely used throughout the country. Criterion-referenced tests are types of measurement devices that are becoming equally, if not more, popular than the more traditional norm-referenced tests.

TESTING

USE OF SPECIFIC INSTRUMENTS OR TECHNIQUES IN THE PROCESS
OF OBSERVING AND RECORDING INFORMATION.

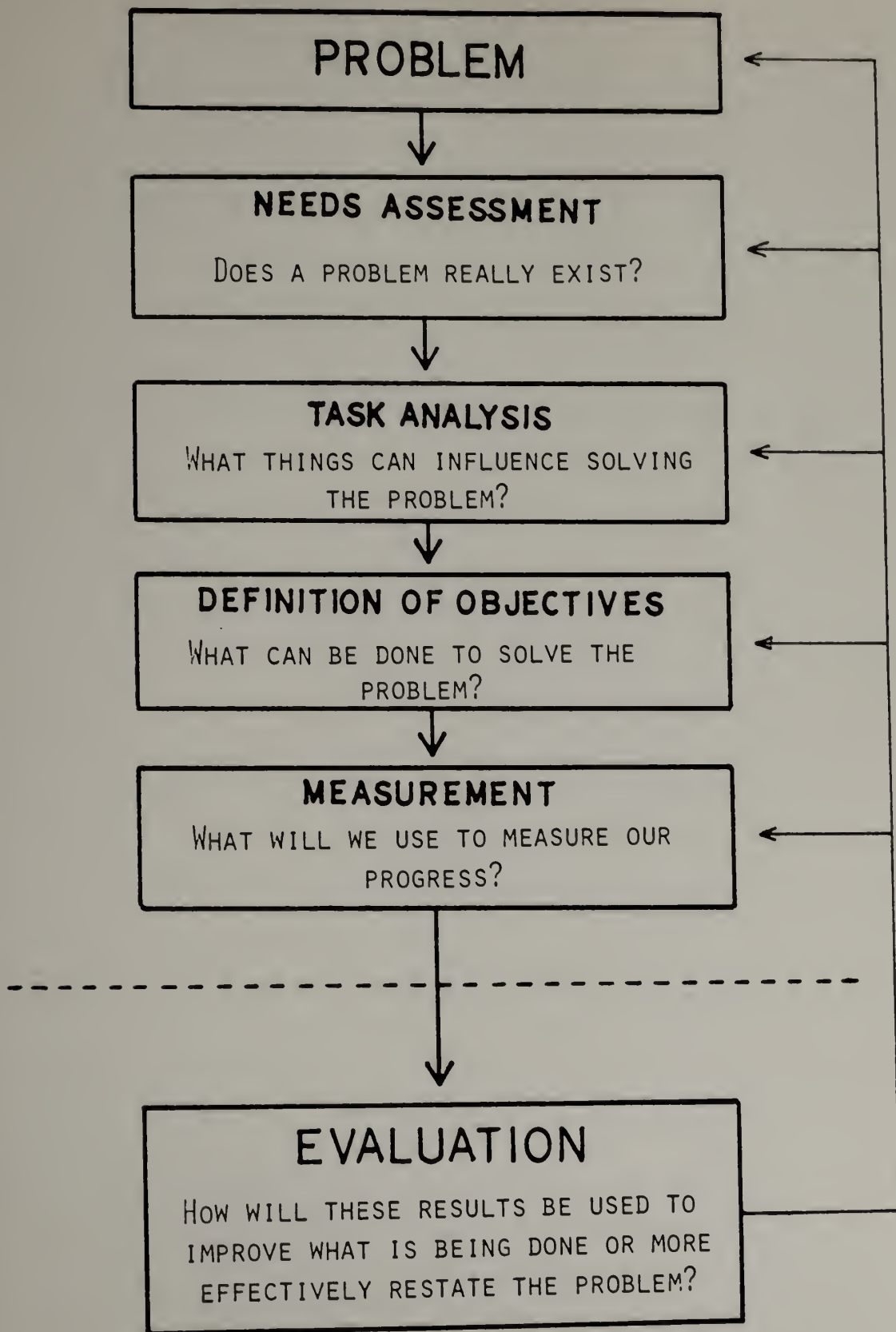
EXAMPLES: SCHOLASTIC APTITUDE TESTS
COMPREHENSIVE TESTS OF BASIC SKILLS
METROPOLITAN ACHIEVEMENT TESTS

GENERAL TYPES: NORM-REFERENCED TESTS
CRITERION-REFERENCED TESTS

SLIDE 6

Assessment is a process of identifying a problem, gathering data to define it and defining objectives to solve it. This part of the process is applied to the measurement of relevant data to determine whether or not the objectives are achieved. Evaluation involves the use of all the steps in this process to provide necessary background data for decision making. Results of the evaluation may affect one or all of the steps in the assessment process. Whether dealing with individual or program evaluation, testing alone is not enough. Test data must be used in concert with other information and must be specific to the problem being assessed before it can be used for effective decision making. Too often "evaluation" of school programs or the system is considered accomplished with the administration of a norm-referenced test. Resulting local averages are compared with national averages as a legitimate index of school or system effectiveness. This is not evaluation in that other relevant information have not been considered in the process. In addition, averages alone are not very sensitive or reliable indices of group performance characteristics.

THE ASSESSMENT PROCESS



SLIDE 7

Evaluation is usually the end product of the assessment process. There are two types of evaluation: Formative and Summative. Formative evaluation involves gathering information on an instructional activity while it is being implemented. The intent is to provide continual feedback to instructional planners, developers and implementors for improvement of instruction while the program or activity is in progress. Summative evaluation is more concerned with the result of an instructional activity. Information is gathered either during or after the implementation of an instructional activity in an effort to judge its overall value. This type of evaluation is frequently used to compare one instructional approach or program with another. Formative evaluators tend to work more closely with instructional planners, developers and implementors. Summative evaluators tend to work more with school administrators or decision-makers. Whatever type of evaluation is used, it should be part of the total instructional development process and not considered only when information is needed to support a program or policy decision.

FORMATIVE EVALUATION

GATHERING INFORMATION DURING THE COURSE OF AN INSTRUCTIONAL ACTIVITY FOR ON-GOING REVISION AND IMPROVEMENT.

RESULTS ARE IN THE FORM OF CONTINUAL FEEDBACK TO INSTRUCTIONAL PLANNERS, DEVELOPERS, AND IMPLEMENTORS.

SUMMATIVE EVALUATION

GATHERING INFORMATION DURING OR AFTER THE COURSE OF AN INSTRUCTIONAL ACTIVITY TO DETERMINE THE ACTIVITIES' OVERALL EFFECTIVENESS.

RESULTS ARE IN THE FORM OF A FINAL REPORT AT THE CONCLUSION OF AN INSTRUCTIONAL ACTIVITY.

SLIDE 8

Norm-referenced tests are a type of test used to measure pupil performance. These emphasize comparing a child's relative position with others in a defined reference group.

NORM-REFERENCED TESTS

- COMPARATIVE MEASURES
- INTERPRETS TEST SCORES OF INDIVIDUALS BY COMPARING THEM TO TEST SCORES OF OTHER INDIVIDUALS.
- MEASURE THE RELATIVE STANDING OF AN INDIVIDUAL IN A GROUP.
- INDICATES THAT JOHNNY CAN DO MORE THAN SUSIE - DOES NOT ANSWER THE QUESTION OF SPECIFICALLY WHAT JOHNNY CAN OR CANNOT DO.

SLIDE 9

The major purposes of norm-referenced tests are to aid in the selection, placement, classification and guidance of students, and overall curriculum management. The term curriculum management is used to emphasize a more prominent role norm-referenced test data can serve in helping educators decide on curriculum approaches. Though norm-referenced test data can be used in providing information about more specific instructional practices, criterion-referenced test data are more useful for this purpose.

PURPOSES OF NORM-REFERENCED TESTS

SELECTION AND PLACEMENT

DETERMINE WHETHER OR NOT AN INDIVIDUAL IS QUALIFIED FOR A PARTICULAR ACTIVITY.

CLASSIFICATION

IF QUALIFIED, IN WHAT PARTICULAR TYPE OF ACTIVITY WOULD THE INDIVIDUAL BE MOST EFFECTIVE.

GUIDANCE

AN AID TOWARD PROVIDING OBJECTIVE DATA FOR MORE EFFECTIVE CLASSIFICATION OR PLACEMENT.

CURRICULUM MANAGEMENT

USED WITH OTHER RELEVANT DATA, CAN CONTRIBUTE ADDITIONAL INFORMATION TOWARD EFFECTIVE DECISION-MAKING ABOUT CURRICULUM.

SLIDE 10

Research evidence indicates there are two aspects which can influence norm-referenced standardized test results; socio-economic status and parent educational level. It is important to note that these outside influences relate to group data not individual data. It is possible for a child from a low socio-economic environment to perform very well on a norm-referenced standardized achievement test. The problem occurs in dealing with group data. In general, pupils from high socio-economic status communities tend to achieve at higher levels than students from low socio-economic status communities. Related to that, children whose parents have a higher level of education generally do better on norm-referenced standardized tests than those whose parents have a lower level of education. These factors are important to consider when discussing the influence of norm-referenced standardized test data, particularly when one is dealing with a test given in a community with diverse population characteristics. Some schools may reflect populations with high socio-economic characteristics, while others may have children from relatively low socio-economic areas. These differences in community make-up may have a decided effect on group test results which will present some problem in the interpretation of town-wide or district data.

OUTSIDE INFLUENCES

PUPILS FROM HIGH SOCIO-ECONOMIC STATUS NEIGHBORHOODS ACHIEVE AT A HIGHER LEVEL THAN PUPILS FROM LOW SOCIO-ECONOMIC STATUS NEIGHBORHOODS.

LEVEL OF EDUCATION COMPLETED BY PARENTS HAS A CLOSE RELATIONSHIP TO PUPIL'S ACADEMIC ACHIEVEMENT.

SLIDE 11

Though the emphasis in this staff development program is on norm-referenced test data, criterion-referenced measurement is becoming very popular -- particularly in the area of program assessment. Consequently, the subject is dealt with briefly - with the intent to discuss the major differences between the two types of measurement approaches.

CRITERION REFERENCED TESTS

DELIBERATELY CONSTRUCTED TO YIELD MEASUREMENTS THAT ARE DIRECTLY INTERPRETABLE IN TERMS OF SPECIFIED PERFORMANCE STANDARDS.

IDENTIFY SPECIFIC SKILL DEFICIENCIES OF INDIVIDUAL LEARNERS.

TEST ITEMS ARE TIED TO SPECIFIC OBJECTIVES AND CAN BE GROUPED UNDER SPECIFIED SKILL LEVELS.

EMPHASIS IS ON HOW WELL PUPILS RESPOND TO ITEMS THAT REPRESENT SKILL AREAS RATHER THAN ON HOW PUPILS COMPARE WITH OTHER PUPILS.

SLIDE 12

Though there are other differences between norm-referenced and criterion-referenced tests, these are listed as major ones. Where a norm-referenced test is generally oriented towards group curriculum in one or more areas, a criterion-referenced test is frequently oriented more towards individualized instruction. Where the norm-referenced test describes a comparative level of knowledge since the results allow a child's score to be compared to a reference group, the criterion-referenced test describes a child's specific level of knowledge as it relates to the skills measured by the test. Where a norm-referenced test references a score to a norm group, the criterion-referenced test references results to some pre-specified criteria. Where a norm-referenced test is usually timed, a criterion-referenced test is not necessarily timed and is usually not timed. Where a norm-referenced test is a general survey of skills, a criterion-referenced test is a comprehensive examination of a discipline -- a much more in-depth examination of a skill area. The results of a norm-referenced test are scores that are based on measures of central tendency, while criterion-referenced test results are indications of mastery or non-mastery of skills based on a child's performance on items representative of those skills. Both types of measurement address similar areas. The criterion-referenced test is usually more suitable for program evaluation because the results can be used to directly assess the performance of children in specific skill areas, and the test itself is composed of items that are designed to be sensitive to an instructional process.

SOME BASIC DIFFERENCES

NRT

CRT

ORIENTED TOWARD GROUP
CURRICULUM

ORIENTED TOWARD INDIVIDUALIZED
INSTRUCTION

DESCRIBES A COMPARATIVE
LEVEL OF KNOWLEDGE AS
DEFINED BY THE TEST

DESCRIBES A SPECIFIC LEVEL OF
KNOWLEDGE AS DEFINED BY THE
TEST

REFERENCES SCORES TO A NORM

REFERENCES SCORES TO CRITERIA

TIMED

NOT NECESSARILY TIMED

SURVEYS GENERAL SKILL AREAS

COMPREHENSIVE EXAMINATION OF
A DISCIPLINE

RESULTS ARE SCORES BASED
ON MEASURES OF CENTRAL
TENDENCY

RESULTS ARE INDICATIONS OF
MASTERY OR NON-MASTERY OF
SKILLS BEING MEASURED

ADDRESS AREAS OF:

ADDRESS AREAS OF:

GUIDANCE
PLACEMENT
RESEARCH
CURRICULUM MANAGEMENT
RESOURCE ALLOCATION

PLACEMENT
INSTRUCTIONAL MANAGEMENT
RESOURCE ALLOCATION
PROGRAM EVALUATION

SLIDE 13

The most common explanation for the difference between norm-referenced measurement and criterion-referenced measurement lies in how results are interpreted. However, these differences are largely due to item construction. Items for norm-referenced tests are designed to discriminate between pupils who do well and those who do not perform well on the test as a whole. The difficulty level of norm-referenced test items is a very important consideration and must correspond to and aim at the group on which the test was originally standardized. Both the discrimination and difficulty level indices are essential in order for norm-referenced tests to rank-order or compare individuals with some defined reference group. These traditional considerations of item discrimination and difficulty level indices have either less or different significance in the criterion-referenced testing environment. The intent of criterion-referenced tests is not necessarily to rank-order or compare individuals, but to determine the level of skill proficiency children have at a given time and be sensitive to the impact instruction may have in eliminating observable skill deficiencies.

TEST CONSTRUCTION DIFFERENCES

NORM-REFERENCED TESTS

ITEMS ARE DESIGNED TO DISCRIMINATE BETWEEN THOSE WHO HAVE HIGH SCORES ON A TEST AND THOSE WHO HAVE LOW SCORES ON A TEST.

ITEMS MISSED BY MOST PUPILS ARE DISCARDED - AS ARE ITEMS PASSED BY MOST PUPILS. THE DIFFICULTY LEVEL OF THE ITEMS SHOULD REFLECT ABOUT A 50% PASS RATE AND MUST RELATE TO THE GROUP ON WHICH THE TEST WAS NORMED.

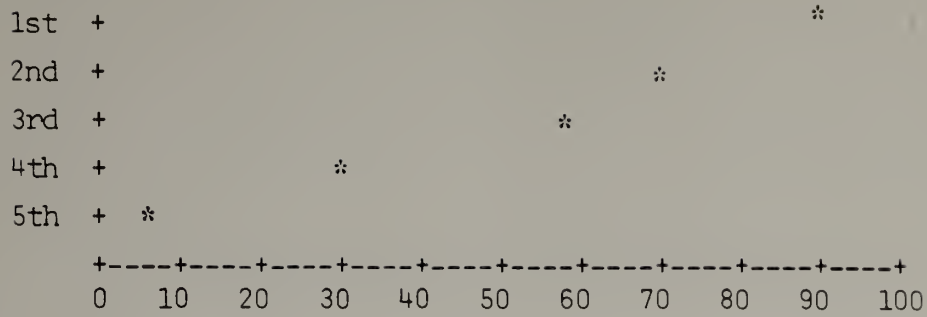
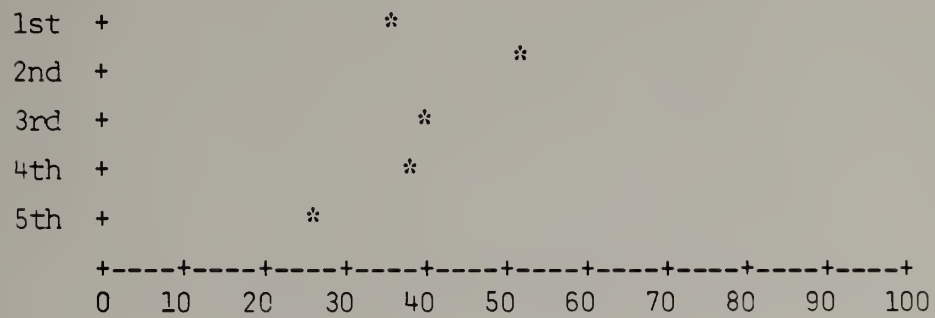
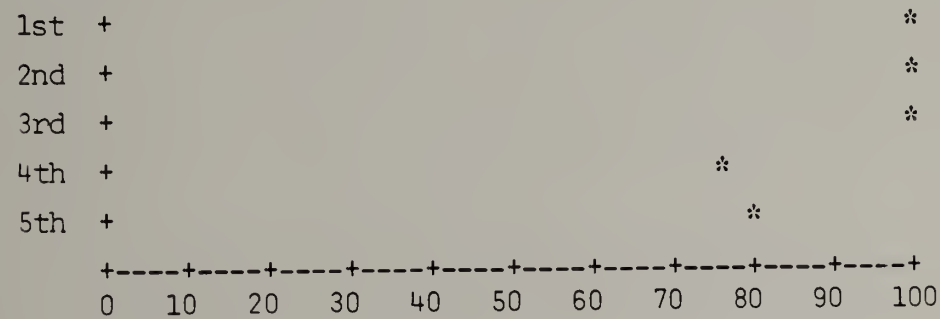
CRITERION-REFERENCED TESTS

ITEMS ARE SELECTED BASED ON THEIR RELEVANCE TO INSTRUCTION - NOT BY THEIR ABILITY TO DISCRIMINATE BETWEEN HIGH AND LOW SCORING PUPILS.

THE DIFFICULTY LEVEL IS NOT RESTRICTED TO HOW A GROUP PERFORMS ON A SAMPLE OF ITEMS AT A PARTICULAR TIME BUT IS A FUNCTION OF THE LEVEL OF MASTERY STATED IN THE PERFORMANCE OBJECTIVES.

SLIDE 14

Different items are selected for different purposes in both norm-referenced and criterion-referenced tests. This transparency shows the performance characteristics of a group of pupils on three items. The total group is divided into fifths according to their total test performance. The percentage of pupils in each scoring fifth are plotted. Item 21 is a good candidate for a norm-referenced test in that pupils who score well on the test as a whole are also scoring well on this item. The percentage of pupils responding correctly to this item decrease as the number of pupils who score lower on the total test increases. Item 22 and 23 would not be good examples of typical norm-referenced test items. Here there is poor discrimination between the pupils grouped by level of overall test performance, and the difficulty level is either too low or too high. However, the characteristics shown by item 22 and 23 would be desirable in a criterion-referenced test if both items measured the same skill and the pattern shown in #22 was apparent before instruction and the pattern shown in item #23 was in evidence after instruction.

Item 21Item 22Item 23

SLIDE 15

Tests having norms may not necessarily be standardized, and a test with well standardized procedures may not necessarily have norms. Standardized tests are not necessarily norm-referenced tests, though these two terms are often considered synonymous. Standardization of procedure is an important concept in the development and administration of norm-referenced tests because it contributes significantly to the validity of the instrument. Procedure standardization may also be an important consideration with a criterion-referenced test or an observational technique. If a criterion-referenced test or observational technique is used for evaluation purposes, the evaluator will want to control and standardize the conditions under which the test or technique is applied.

STANDARDIZED TESTS

- A STANDARDIZED TEST IS ONE IN WHICH THE PROCEDURES FOR ADMINISTERING AND SCORING THE TEST HAVE BEEN ESTABLISHED AND MUST BE FOLLOWED EACH TIME IT IS GIVEN.
- A TEST MAY HAVE NORMS AND NOT BE STANDARDIZED.
- A TEST MAY BE WELL STANDARDIZED AND NOT HAVE NORMS.

EXAMPLES OF
NORM-REFERENCED TESTS

SLIDE 17

Norm-referenced tests emphasize comparative information. The following is an example of a type of norm-referenced test report for a child. The scores are not explained in any great detail at this point since they will be covered later in the training program. The important consideration in norm-referenced data is that the primary information is expressed as scores. Most test publishers also provide item analysis information which is also shown later in this program.

NORM-REFERENCED TEST RESULTS

NAME: SHIRLEY JONES

GRADE: 6.0

	<u>RAW</u> <u>SCORE</u>	<u>GRADE</u> <u>EQUIV</u> <u>SCORE</u>	<u>STANDARD</u> <u>SCORE</u>	<u>NATL</u> <u>%ILE</u>	<u>LOCAL</u> <u>%ILE</u>
READING VOCABULARY	35	8.3	541	80	59
READING COMPREHENSION	32	6.9	504	56	34
READING TOTAL	67	7.5	510	68	44
SPELLING	39	5.8	477	48	44
LANGUAGE MECHANICS	15	7.2	507	56	42
LANGUAGE EXPRESSION	27	8.7	550	72	46
LANGUAGE TOTAL	81	7.2	498	60	45
MATH COMPUTATION	34	5.9	442	47	33
MATH CONCEPTS	16	6.1	454	50	26
MATH APPLICATIONS	9	3.9	385	20	5
MATH TOTAL	59	5.5	427	38	19
TOTAL BATTERY	207	6.4	459	55	32

SLIDE 18

The intelligence or aptitude test is another type of norm-referenced test. They are probably the most misunderstood and misused test on the market. Some of the problems associated with these tests are shown as well as ways in which these tests can serve the academic community.

INTELLIGENCE/APTITUDE

DICTIONARY DEFINITIONS

INTELLIGENCE TESTS: DESIGNED TO MEASURE THE CAPACITY TO LEARN APART FROM ACTUAL ACHIEVEMENT.

APTITUDE TESTS: DESIGNED TO PREDICT AN INDIVIDUAL'S ABILITY TO LEARN CERTAIN SKILLS.

- BOTH ARE SIMILAR IN PURPOSE.
- NEITHER ARE DESIGNED AS COMPREHENSIVE MEASURES OF INNATE INTELLIGENCE.
- BOTH MEASURE "CAPACITY TO LEARN" WITHIN THE LIMITS OF THE SPECIFIC AREAS TESTED.
- BOTH MEASURE ABILITY TO PERFORM CERTAIN TYPES OF MENTAL TASKS SUCH AS:
 - VOCABULARY
 - ANALOGIES
 - SEQUENCES
 - MEMORY
- BOTH CAN BE USED TO PREDICT AREAS OF ACADEMIC STRENGTH AND WEAKNESS.

EXAMPLES OF
CRITERION-REFERENCED TESTS

SLIDE 20

Results from criterion-referenced tests usually reflect pupil performance on items representing specific skills. The example shown is a report for a pupil who has taken a primary level math test. Each skill shown is measured by several items and the results are keyed to the following explanations:

"M" = Mastery - All items for a given skill are answered correctly.

"R" = Review - At least 50% of the items for a given skill are answered correctly.

"L" = Needs to Learn - Less than 50% of the items for a given skill are answered correctly.

"O" = Omit - All items for a given skill were omitted.

PUPIL REPORT - OCTOBER, 1976

SCHOOL : RIDGEWOOD

TEACHER: BRADLEY

LEVEL : 2

STUDENT: KEVIN ADAMS

OBJECTIVE CODE	OBJECTIVE DESCRIPTION	SKILL LEVEL
SE201	CREATES SEMI-CONCRETE EQUIV. SETS (0-9)	M
SE202	CREATES SEMI-CONC. PICTURE OF SETS (0-18)	M
NU202	READS & EXPLAINS NON-EQUIV. STATEMENTS	M
NU203	USES ORDINAL NUMBERS THROUGH NINTH	M
NU204	ORDERS NUMBERS 0-99	M
PV201	WRITES NUMERALS 1-99 FROM MEMORY	M
PV202	USES & READS INEQUALITIES OF 2-DIGIT NOS.	M
AS203	ADDS & SUBTRACTS (0-9) HORIZ & VERT	M
AW202	FINDS MISSING ADDEND IN 1ST PLACE (0-9)	M
ME202	USES CENTIMETER & INCH TO MEASURE	M
EF203	CONSTRUCTS REGION WITH 2 EQUAL PARTS	M
EF204	ASSOCIATES 1/2 WITH ONE-HALF, ETC.	M
GE202	IDENTIFIES BASIC SHAPES	M
GE204	IDENTIFIES ANGLES OF SAME SIZE	M
MO201	SHOWS RLTN SHP PENNY, NICKEL, DIME, QTR.	M
MO202	SHOWS RLTN SHP PENNY, NICK, DIME, QTR, 1/2, DULL.	M
TI201	TELLS TIME TO HALF-HOUR	M
TI202	TELLS TIME TO NEAREST QUARTER-HOUR	M
TI204	SAYS DAYS OF WEEK IN ORDER	M
NU201	WRITES (0-9) NON-EQUIV NOS. WITH > AND <	R
AS205	ADDS & SUBTRACTS (10-18) HORIZ & VERT	R
AW201	FINDS MISSING ADDEND IN 2ND PLACE (0-9)	R
GE201	LABELS PTS. IN LINE & NAMES LINE SEGS.	K
GE203	IDENTIFIES ANGLES & CORNERS & APPLIES	R
PS201	SOLVES STORY PROBLEMS USING ADDITION & SUBTR.	R
SE203	IDENTIFIES LESSER, GREATER & EQUIV. SETS	L
SE204	NAMES CARDINAL NUMBERS OF A SET	L
NU205	COUNTS TO 500 BY 2, 5 AND 10	L
PV203	GROUPS & NAMES HUNDREDS, TENS, ONES IN A NO.	L
PV205	WRITES 3-DIG. NOS. FROM PICTURES OF OBJECTS	L
PV207	IDENTIFIES GREATER/LESSER OF 3-DIGIT NOS.	L
AS204	ADDS & SUBTRACTS (10-18) USING OBJECTS	L
EF201	IDENTIFIES 1/4 OF GIVEN SET	L
EF202	IDENTIFIES 1/3 OF GIVEN SET	L
PV204	NAMES NO. OF HUNDREDS, TENS, ONES IN 3-DIG. NO.	O
PV206	COMPLETES SEQUENTIAL PATTERN OF 3-DIGIT NOS.	O
PV208	COUNTS WITH 3-DIGIT NUMBERS	O
PV209	WRITES EXPANDED NUMBERS (10-999)	O
AS201	EXPLAINS "SUM" & "DIFFERENCE"	O
AS207	ADDS & SUBTRACTS 2-DIG. NOS. W/OUT RENAMING	O
ME201	USES LINEAR MEASURE (IN., FT., YDS., HALF-IN.)	O
TI203	TELLS TIME TO NEAREST 5 MINS.	O
TI206	TELLS TIME TO NEAREST MINUTE	O
TI205	SAYS MONTHS OF YEAR IN ORDER	O
NT201	IDENTIFIES NUMBER AS ODD OR EVEN	O
NT202	PREDICTS SUM OF 2 NOS. AS ODD OR EVEN	O

SLIDE 21

Advantages of criterion-referenced approaches to measurement are the ability to both diagnose skill deficiencies and direct the teacher and/or pupil into materials to correct those deficiencies. An example of this type of measurement is the Prescriptive Reading Inventory.* This report is for an individual showing the objectives tested and mastered (+), not mastered (-), or needing review (R). The level of mastery is determined by how well this pupil performed on the items measuring each of these skills.

* CTB/McGraw-Hill, Monterey, California.

+ = MASTERY OF OBJECTIVE, R = REVIEW RECOMMENDED, - = NONMASTERY OF OBJECTIVE

- PHONIC ANALYSIS
- 9. Silent Letters R
 - 13. Variant Vowel Sounds: Digraph, Diphthong -
 - 14. Phonetic Parts: Variant Sounds -
 - 15. Phonetic Parts: Blending R

- STRUCTURAL ANALYSIS
- 22. Pronouns: Referent R
 - 25. Compounds: Forming +
 - 30. Sentence Building: Phrase Selection -
 - 31. Phrase Information +
 - 32. Affixes: Identifying Prefixes, Suffixes +
 - 33. Affixes: Building Words -
 - 34. Defining Affixed Words R
 - 37. Punctuation: Exclamation Point -

TRANSLATION

- 45. Meaning of Related Words in Context +
- 46. Most Precise Word in Context -
- 48. Word Definition in Context -
- 49. Word Definition in Isolation +
- 51. Multi-meaning Words and Synonyms +
- 52. Synonyms: Selection +
- 53. Antonyms: Selection -
- 54. Homonym Pairs: Selection +

- LITERAL COMPREHENSION
- 57. Event Sequence -
 - 58. Story Setting +
 - 59. Story Detail: Recall or Descriptive Words +
 - 60. Story Detail: Recall by Parts R
 - 61. Story Detail: Identify True Statement +

INTERPRETIVE COMPREHENSION

- 62. Cause or Effect -
- 63. Inference +
- 64. Conclusion: Formation -
- 66. Predicting Future Action +
- 67. Main Idea: Summary, Title or Theme +
- 70. Character Analysis: Descriptive Words, Traits R
- 71. Descriptive Words or Phrases +
- 72. Sensory Imagery -
- 73. Idioms or Figures of Speech -
- 75. Simile -
- 76. Metaphor -
- 77. Mood -
- 78. Time Span and Period -

CRITICAL COMPREHENSION

- 80. Literary Forms: Fable -
- 83. Reality and Fantasy R
- 84. Reality and Fantasy Possibility +
- 89. Author Purpose -

SLIDE 22

If problem areas for the individual can be identified, one can also group those with similar skill deficiencies together. This process of grouping by similar skill deficiencies is shown on this report where pupils with the same skill problems are noted with an asterisk. In this example, phonic analysis, interpretive comprehension and critical comprehension appear to be skill areas needing attention.

TEACHER CORNING 31 GRADE 3 PROCESS NUMBER 2010-101
 SCHOOL MILLER 3 CITY SOUTH FALLS 1 DATE OF TESTING 09/72
 STATE CA RUN DATE 10/03/72

PRESCRIPTIVE READING INVENTORY
 CLASS GROUPING REPORT
 BLUE BOOK (C)

NAMES	PHONIC ANALYSIS		STRUC ANALYSIS	TRANSLATION	LITERAL COMP		INTERP COMP		CRITICAL COMP
	I	I			I	I	I	I	
BRISCOE, NANCY									
BROWNE, FRANKLIN									
CALABRESE, TONY									
DAVIDSON, MAL									
DAVIS, ERNESTINE									
ELDER, LYNN									
FELICIANO, MIGUEL									
HARRIS, JEAN									
MINOR, MORRIS									
WOOD, JENNIFER									
TOTAL IN GROUP	7	6	5	4	3	8	7		

NOTE - IF A COLUMN IS EMPTY, NO STUDENT NEEDS INSTRUCTION IN THAT CATEGORY

PHONIC ANALYSIS	NO. STUDENTS FOR OBJECTIVE	LITERAL COMPREHENSION	NO. STUDENTS FOR OBJECTIVE
9 SILENT LETTERS.....	4	57 EVENT SEQUENCE.....	3
13 VARIANT VOWEL SOUNDS-DIGRAPH, DIPHTHONG.....	7	58 STORY SETTING.....	4
14 PHONETIC PARTS-VARIANT SOUNDS.....	7	59 STORY DETAIL-RECALL OR DESCR. WORDS.....	4
15 PHONETIC PARTS-BLENDING.....	5	61 STORY DETAIL-IDENTIFY TRUE STATEMENTS.....	2
STRUCTURAL ANALYSIS		INTERPRETIVE COMPREHENSION I	
22 PRONOUNS-REFERENT.....	4	62 CAUSE OR EFFECT.....	2
25 COMPOUNDS-FORMING.....	1	63 INFERENCE.....	2
30 SENTENCE BUILDING-PHASE SELECTION.....	3	64 CONCLUSION-FORMATION.....	3
31 PHRASE INFORMATION.....	6	66 PREDICTING FUTURE ACTION.....	2
32 AFFIXES-IDENTIFYING PREFIXES,SUFFIXES.....	4	70 CHARACTER ANALYSIS-DESCR.WORDS,TRAITS.....	3
33 AFFIXES-BUILDING WORDS.....	6	INTERPRETIVE COMPREHENSION II	
34 DEFINING AFFIXED WORDS.....	5	71 DESCRIPTIVE WORDS AND PHRASES.....	4
37 PUNCTUATION-EXCLAMATION POINT.....	6	72 SENSORY IMAGERY.....	3
TRANSLATION		73 IDIOMS OR FIGURES OF SPEECH.....	8
45 MEANING OF RELATED WORDS IN CONTEXT.....	2	75 SIMILE.....	7
46 MOST PRECISE WORD IN CONTEXT.....	5	76 METAPHOR.....	8
48 WORD DEFINITION IN CONTEXT.....	5	77 MOOD.....	5
49 WORD DEFINITION IN ISOLATION.....	3	78 TIME SPAN AND PERIOD.....	6
51 MULTI-MEANING WORDS AND SYNONYMS.....	4	CRITICAL COMPREHENSION	
52 SYNONYMS-SELECTION.....	3	80 LITERARY FORMS-FABLE.....	4
53 ANTONYMS-SELECTION.....	5	83 REALITY AND FANTASY.....	5
		84 REALITY AND FANTASY-POSSIBILITY.....	6
		89 AUTHOR PURPOSE.....	7

From the Prescriptive Reading Inventory Interpretive Handbook. Reproduced by permission of the publisher, CTB/McGraw-Hill, Monterey, CA 93940. Copyright © 1972 by McGraw-Hill, Inc. All Rights Reserved. Printed in the U.S.A.

SLIDE 23

An advantage of criterion-referenced tests is diagnosing specific skill deficiencies making it possible to prescribe ways to correct the deficiencies. This report is for Don Bates whose teacher has identified as working in Around Green Hills, a textbook published by American Book Company. The pages in this text are given where he and his teacher may turn to seek help in correcting tested skill deficiencies.

P R E S C R I P T I V E R E A D I N G I N V E N T O R Y
I N D I V I D U A L S T U D Y G U I D E
R E D B O O K

TEACHER FORBES GRADE 2
SCHOOL MORRISON PROCESS NUMBER 1762
CITY ANY TOWN DATE OF TESTING 09/72
STATE CA RUN DATE 09/15/72

NAME BATES DONALD

TEXTBOOK - AROUND GREEN HILLS
AMERICAN BOOK CO.

IR
1965

O B J E C T I V E

STUDENT EDITION *
REFERENCE

TEACHER EDITION
REFERENCE

WORKBOOK
REFERENCE

O B J E C T I V E	STUDENT EDITION * REFERENCE	TEACHER EDITION REFERENCE	WORKBOOK REFERENCE
RECOGNITION OF SOUND AND SYMBOL	I	195G	33,74
3 VOWEL SOUNDS-UNLIKE	I		
PHONIC ANALYSIS	I		
5 CONSONANT SUBSTITUTION-INIT. AND FINAL	I		
6 CONSONANT SUBSTITUTION-FINAL	I	96,120G,127G	26,64,71
7 SYLLABLES-NUMBER	I	121	36
STRUCTURAL ANALYSIS	I	141G,227-228W	45,65
16 INFLECTED WORDS-SINGULAR/PLURAL	I	130W,180W,239W	31,40
17 INFLECTED WORDS(ENDINGS)AND AFFIXES	I	49G,62G,91,92G	5,10,24,27,45,65
19 ADJECTIVES-POS.,COMP.,SUPERL.	I		
21 PRONOUNS	I		
23 CONTRACTIONS-WORD PAIRS/VERB PHRASES	I		
29 SENTENCE BUILDING-SUBJECT,PREDICATE	I		
TRANSLATION	I		
38 LIKE AND UNLIKE ENTITIES-WORD DEF.	I	227-228G,239	1,11,16,22,40,85,86,87
39 LIKE AND UNLIKE ENTITIES-SYNONYMS	I		
42 USE OF CONTEXT-SENTENCE COMPLETION	I		
43 HOMONYMS IN CONTEXT	I		
44 SENTENCE SENSE	I		
LITERAL COMPREHENSION	I		
59 STORY DETAIL-RECALL OR DESCR. WORDS	I	139,172,193,214	11,16,19,22,35,44,56,60,72,82
INTERPRETIVE COMPREHENSION	I		
64 CONCLUSION-FORMATION	I	174G,209,226	6,56,62,63,79,87,90
69 CHARACTER ANALYSIS-MOTIVE OR CAUSE	I		
CRITICAL COMPREHENSION	I		
79 PROBLEM SOLUTION	I		

SLIDE 24

Another example of a published criterion-referenced test is the Diagnostic Mathematics Inventory.* This example from the DMI emphasizes the need to identify skills the teacher wishes to measure prior to giving the test. In this way, each testing situation is specified by the teacher rather than dictated by the test. The types of skills covered in each test level are described and the specific item numbers measuring these skills are given. Consequently, a teacher has the option of defining what skills are to be measured and test either individuals or groups on just those skills.

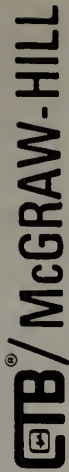
* CTB/McGraw-Hill, Monterey, California

LEVEL C/BLUE

- 1 Addition of Whole Numbers Without Regrouping: 12, 15, 17, 18
- 2 Addition of Whole Numbers With Regrouping: 21, 22, 23, 25
- 3 Subtraction of Whole Numbers Without Regrouping: 26, 29, 31
- 4 Subtraction of Whole Numbers With Regrouping: 34, 36, 37
- 5 Multiplication of Whole Numbers (Preoperational): 39, 40, 41
- 6 Multiplication of Whole Numbers: 44, 45, 46, 47, 48
- 7 Division of Whole Numbers: 56, 57, 58, 59, 60
- 8 Fractions: 68, 69, 70, 71, 88
- 9 Commutative, Associative, and Distributive Properties: 137, 139, 141, 143, 145
- 10 Identity Element and Inverse Relations: 148, 150, 151, 152, 155
- 11 Rounded Numbers and Estimation: 158, 159, 160
- 12 Sequences: 164, 165, 166
- 13 Missing Addends and Factors: 173, 176, 178
- 14 Inequalities and Number Theory: 2, 183, 184, 186, 188
- 15 Metric Geometry: 198, 200, 204
- 16 Linear Measure: 209, 210, 211, 212
- 17 Money: 217, 218, 219, 220
- 18 Weight, Liquid, Dozen: 222, 226, 229
- 19 Temperature and Time: 230, 232, 233
- 20 Points, Segments, Lines, Rays: 238, 240, 241, 278
- 21 Plane Figures: 244, 245, 254, 257, 259, 262, 267
- 22 Place Value: 279, 280, 281, 282

SLIDE 25 - 27

The following are reports generated from DMI data. The first (Individual Diagnostic Report) is for an individual pupil with indications of skills mastered (+) or not mastered (-). The second report (Objective Mastery Report) shows the same type of information but the data are organized by class or instructional grouping for better classroom instructional management. The third report (Class Pre-Mastery Analysis) shows an asterisk for each pupil demonstrating common types of math errors in their performance on the test.



ADDITION OF WHOLE NUMBERS NO REGROUPING
 003 NUMBER LINE (POINT)
 012 NUMBER LINE (ADDITION)
 019 ADDITION 5-DIGIT

ADDITION OF WHOLE NUMBERS, REGROUPING
 021 ADDITION 2-DIGIT (RG)
 022 ADDITION 3-DIGIT (RG)
 023 ADDITION 4-DIGIT (RG)
 024 ADDITION 5-DIGIT (RG)
 025 ADDITION COLUMN

SUBTRACTION OF WHOLE NUMBERS NO REGROUPING
 026 NUMBER LINE (SUBTRACTION)
 032 SUBTRACTION 1D - 1D = 1D

SUBTRACTION OF WHOLE NUMBERS, REGROUPING
 034 SUBTRACTION 2D - 2D = 2D (RG)
 036 SUBTRACTION 3D - 3D = 3D (RG)
 037 SUBTRACTION 4D - 4D = 4D (RG)
 038 SUBTRACTION 5D - 5D = 5D (RG)

MULTIPLICATION OF WHOLE NUMBERS NO REGROUPING
 039 NUMBER LINE (MULTIPLICATION)
 040 MULTIPLICATION: REPEATED ADDITION
 041 MULTIPLICATION: ROWS/COLUMNS
 045 MULTIPLICATION: BASIC FACTS
 054 MULTIPLICATION: POWERS OF TEN

MULTIPLICATION WHOLE NUMBERS REGROUPING
 046 MULTIPLICATION 1D X 2D
 047 MULTIPLICATION 1D X 3D
 049 MULTIPLICATION 1D X 4D
 050 MULTIPLICATION 2D X 2D

DIVISION OF WHOLE NUMBERS
 056 NUMBER LINE (DIVISION)
 058 DIVISION, BASIC FACTS, NO REMAINDER
 059 DIVISION 3D ÷ 1D, NO REMAINDER
 060 DIVISION 2D ÷ 1D, WITH REMAINDER
 067 DIVISION, POWERS OF TEN

ADDITION OF DECIMAL NUMBERS NO REGROUPING
 101 DECIMAL FRACTIONS ONE PLACE
 102 DECIMAL FRACTIONS TWO PLACES
 103 DECIMAL FRACTIONS THREE PLACES

ADDITION OF DECIMAL NUMBERS REGROUPING
 104 DECIMAL FRACTIONS ONE PLACE RG
 105 DECIMAL FRACTIONS TWO PLACES RG
 106 DECIMAL FRACTIONS THREE PLACES RG

SUBTRACTION DECIMAL NUMBERS NO REGROUPING
 107 DECIMAL FRACTIONS ONE-DIGIT
 108 DECIMAL FRACTIONS TWO-DIGIT

SUBTRACTION OF DECIMAL NUMBERS REGROUPING
 109 DECIMAL FRACTIONS ONE-DIGIT RG
 110 DECIMAL FRACTIONS TWO PLACES
 111 DECIMAL FRACTIONS THREE-DIGIT RG

COMMUTATIVE, ASSOCIATIVE, DISTRIBUTIVE PROPERTIES
 137 MISSING ADDEND, COMMUTATIVITY
 139 MISSING FACTOR, COMMUTATIVITY
 141 MISSING ADDEND, ASSOCIATIVITY
 143 MISSING FACTOR, ASSOCIATIVITY
 145 MISSING NUMBER, DISTRIBUTIVE

IDENTIFY ELEMENT, INVERSE RELATIONSHIPS
 148 MISSING ADDEND OR SUM, IDENTITY
 150 MISSING FACTOR, IDENTITY
 153 MISSING ADDEND, ADD/SUBT.

NUMBER SEQUENCES
 164 LETTERS
 165 ADDITION - WHOLE NUMBERS
 166 SUBTRACTION - WHOLE NUMBERS

INEQUALITIES, ODDS, MULTIPLES
 185 WHOLE NUMBERS
 187 EVEN AND ODD NUMBERS
 188 MULTIPLES

METRIC GEOMETRY
 200 AREA (WHOLE NUMBERS)
 201 VOLUME (WHOLE NUMBERS)
 204 GRAPHS

LINEAR MEASURE
 209 LINEAR (NON-STANDARD UNITS)
 210 LINEAR (M., CM., MM.)

MONEY
 217 MONEY (CONCEPT)
 218 MONEY (ADDITION)
 219 MONEY (SUBTRACTION)
 220 MONEY (MULTIPLICATION)

WEIGHT AND LIQUID MEASURE
 223 WEIGHT (1LBS. OZS.)
 226 LIQUID MEASURES (CONCEPT)
 227 LIQUID MEASURES (ADD.)
 228 LIQUID MEASURES (SUBT.)

TEMPERATURE AND TIME
 230 TEMPERATURE
 232 CLOCK (CONCEPT)
 233 CALENDAR (DATES)

SEGMENTS, LINES, RAYS
 239 LINE SEGMENTS
 240 PARALLEL LINES
 241 PERPENDICULAR LINES
 243 RAY

PLACE VALUE
 279 0-999
 282 1,000 - 99,999

EXPANDED NOTATION
 280 EXPANDED NOTATION (3-DIGIT)
 281 EXPANDED NOTATION (ADDITION)

OBJECTIVES MASTERY REPORT

TEACHER COWDON J
 SCHOOL SAN ANDREAS
 CITY REEF CITY
 STATE CA

GRADE 5.1
 BATCH 500J-004
 DATE OF TESTING 10/75
 RUN DATE 10/30/75

TEST LEVEL DIAGNOSTIC MATHEMATICS INVENTORY
 D/ORANGE

A B B C C C D D D F F G G H H J J L M M P P R S S T W W
 M A E L A E O A I I A O R A R A U A E O A E O A E O H T O I Y
 E R A O R C L V E L H L O R E R N S R H C S E R U T B A E M L M
 S N R O N I L I B L E E S D E R T O E N C E O R R L E E N V B D A
 E D M E L I S L O Y Y T N N I N M S Y N S E S R R N E L O N
 J S Y N E N E S I O L O S T O N I N
 O A D A T A P B M R S D C R N M M S N S N S N R
 H R M O N O B N D M A A A O P A I O A I T M J B O
 N O Y N N M I G A A T R R J N E N N A R L E A P E G S W J E S
 B L I N T B Y I I T D N G T V D A A L H I U T S

	MAS	PCT
NUMBER OF CASES - 32		
ADDITION OF WHOLE NUMBERS WITHOUT REGROUPING		
003 NUMBER LINE (POINT)	94	+
012 NUMBER LINE (ADDITION)	84	+
019 ADDITION 5-DIGIT	63	+
ADDITION OF WHOLE NUMBERS WITH REGROUPING		
021 ADDITION 2-DIGIT (RG)	81	-
022 ADDITION 3-DIGIT (RG)	63	+
023 ADDITION 4-DIGIT (RG)	50	+
024 ADDITION 5-DIGIT (RG)	41	+
025 ADDITION COLUMN	47	+
SUBTRACTION OF WHOLE NUMBERS WITHOUT REGROUPING		
026 NUMBER LINE (SUBTRACTION)	88	+
032 SUBTRACTION 1D - 1D = 1D	84	+
SUBTRACTION OF WHOLE NUMBERS WITH REGROUPING		
034 SUBTRACTION 2D - 2D = 2D (RG)	56	-
036 SUBTRACTION 3D - 3D = 3D (RG)	63	+
037 SUBTRACTION 4D - 4D = 4D (RG)	47	-
038 SUBTRACTION 5D - 5D = 5D (RG)	22	-
MULTIPLICATION OF WHOLE NUMBERS WITHOUT REGROUPING		
039 NUMBER LINE (MULTIPLICATION)	88	+
040 MULTIPLICATION: REPEATED ADDITION	78	+
041 MULTIPLICATION: ROWS/COLUMNS	78	+
045 MULTIPLICATION: BASIC FACTS	63	+
054 MULTIPLICATION: POWERS OF TEN	59	-

Reproduced by permission of the publisher.



Published by CTB/McGraw-Hill, Del Monte Research Park, Monterey, California 93940. Copyright © 1975 by McGraw-Hill, Inc. All Rights Reserved. Printed in the U.S.A.

CLASS PRE-MASTERY ANALYSIS

TEACHER CONDON J
 SCHOOL SAN ANDREAS
 CITY REEF CITY
 STATE CA
 GRADE 5.1
 BATCH 5001-004
 DATE OF TESTING 10/75
 RUN DATE 10/30/75

TEST DIAGNOSTIC MATHEMATICS INVENTORY
 LEVEL D/ORANGE

A B B B C C C D D D F F F G G H H J J J L L M M P P R S S T W W
 M A E L A E D A I I A O R A R A U A E O A A E O A E O H T O I Y
 E R A O R C L V E L H L O R E R N S R H C S E R U T B A E M L M
 S N R O N I L I B L E E S D E R T O E N E O R L E E N V B D A
 E D M E L I S L O Y Y T N N I N M S Y N S E S R R N E L O N
 J S Y N E N E S I O L O S T O N I N
 O A D A T A P B M R S D C R N M M S N S N S N R
 H R M O N B N D M A A A O P A I O A I T M J B O
 N O Y N N M I G A A T R R J N E N N A R L E A P E G S W J E S
 B L I N T B Y I I T D N G T V D A A L H I U T S

	NUMBER OF CASES - 32	N
ADDING WHOLE NUMBERS		
AW1 MISADDED	12	*
AW2 DID NOT CARRY	8	**
AW3 CARRIED WRONG DIGIT	3	*
AW4 MULTIPLIED INSTEAD	1	*
AW5 ADDED LEFT TO RIGHT	3	*
AW6 SUBTRACTED INSTEAD	2	*
SUBTRACTING WHOLE NUMBERS		
SW1 MINUEND FROM SUBTRAHEND	2	*
SW2 BORROWED COLUMN NOT REDUCED	20	**
SW3 DID NOT BORROW FROM ZERO	18	**
SW4 SUBTRAHEND > MINUEND ZERO	13	**
SW5 ADDED INSTEAD	4	*
SW6 CARRIED INSTEAD OF BORROWING	11	**
SW7 MISSUBTRACTED	19	**
MULTIPLYING WHOLE NUMBERS		
MW1 MULTIPLIED WITHIN COLUMN ONLY	14	**
MW2 DID NOT ADD CARRIED DIGIT	10	**
MW3 MISALIGNED PARTIAL PRODUCTS	8	*
MW4 ADDED INSTEAD	4	*
MW5 SUBTRACTED INSTEAD	7	*
MW6 MISMULTIPLIED	19	**
MW7 MISMULTIPLIED (REPEAT ADD)	12	**
MW8 MISMULTIPLIED (MATRIX)	10	**
MW9 RECORDED UNRELATED PRODUCT	3	*



Reproduced by permission of the publisher.

Published by CTB/McGraw-Hill, Del Monte Research Park, Monterey, California 93940 Copyright © 1975 by McGraw-Hill, Inc. All Rights Reserved. Printed in the U.S.A.



SLIDE 28

Teacher-made tests are developed as both criterion-referenced and norm-referenced types of measures. Test items are written to measure pupil performance in particular skill areas, and are designed to be sensitive to what is taught in the classroom. Results of these types of tests are also used in a norm-referenced way when pupils' scores are ranked and used for grading or instructional placement purposes. Some teachers collect data on their tests over several years and compare each new class with previous pupil performance. Whatever the use, there are problems with development and application of teacher-made tests and their results. Some of these problems are discussed below. The intent of this section is to make educators aware and focus their attention on seeking solutions to the problems, not in teaching instructors how to construct classroom tests.

PURPOSE - Before a teacher begins the process of test construction, it is necessary to define what is to be measured and why. By outlining specifically what is to be measured, a teacher can begin to construct items responsive to classroom instructional needs. The type of instrument constructed will depend on the type of information desired. A teacher may wish to construct a different type of test if the intent is to rank order children on the basis of total test performance in a particular skill domain, rather than determine the types of errors children may be making in a specific skill area.

CONSTRUCTION - Tests generally fall into two categories, Essay and Objective tests. There are fairly well established rules published in several texts and pamphlets for the construction of items for both types of tests. Examples of these will be distributed at the time this part of the staff development program is implemented. Item construction should be responsive to the purpose of the information gathering process pre-specified by the teacher. Different construction strategies would be used if the test was to have norm-referenced rather than criterion-referenced implications.

ANALYSIS - The value of the item construction process can be determined through an analysis of results from a trial testing of an instrument. An appropriate way to analyze data from the trial is to examine pupil responses to individual items, referred to as item analysis. Most tests developed by classroom teachers are called multiple choice tests. Each item has several response options and only one is the most desirable or correct answer. Item analysis can help the teacher determine if the item is properly constructed. It is possible the teacher may find some of the response options are confusing to the pupils and not appropriate for determining the level of skill proficiency desired. Item analysis is also used to determine the level of item difficulty and how well the item can separate children who score either high or low on the test as a whole - both of these are important considerations if the test is to have norm-referenced interpretations. An excellent source document for teachers to use in the analysis of teacher-made tests is written by Paul B. Diederich of Educational Testing Service titled - Short-Cut Statistics for Teacher-Made Tests, (1973). This document provides some simple procedures for item analysis and computing basic statistics necessary for the analysis of test results.

INTERPRETATION AND APPLICATION - After the analysis phase, the interpretation and application of results to instructional improvement is an important consideration. If the items were constructed to reveal diagnostic information, then the teacher would not be making effective use of the instrument if scores rather than item response data were the primary or only use of the results.

TEACHER - MADE TESTS

NORM - REFERENCED
USAGE

CRITERION - REFERENCED
USAGE

PROBLEMS

- PURPOSE
- CONSTRUCTION
- ANALYSIS
- INTERPRETATION
- APPLICATION

SLIDE 1

The statistics covered in this section are limited to the most useful terms for instructional interpretation and use of test data. This section of the program deals with defining and discussing both the strengths and weaknesses of these terms used in the instructional environment.

TECHNIQUES

TYPES

- STATISTICS
- SCORES & SCORE DISTRIBUTIONS
- ITEM ANALYSIS

SLIDE 2

The mean, median and mode of a distribution of 21 scores is displayed. Their relative positions are indicated to show that an "average" has three definitions. Any one or all three may be used depending on the nature of the data and what message the user needs.

SCORES

8
 9
 9
 13
 13
 15
 17
 17
 17
 20
 22
 25
 25
 25
 25
 26
 27
 27
 27
 28
 30

- MEAN
 - MEDIAN
 } MODE

$N = 21$

$SUM = 425$

$MEAN = 20.2$

$MEDIAN = 22$

$MODE = 25$

SLIDE 3

The mean, an arithmetic average of the scores in a distribution, is the most commonly used measure of central tendency. The mean differs from the other averages in that all scores in the distribution are used in its calculation.

MEAN

THE MOST COMMON INDICATOR OF CENTRAL TENDENCY - AN ARITHMETIC AVERAGE OF SCORES.

$$\text{MEAN} = \frac{\text{SUM OF SCORES}}{\text{NUMBER OF SCORES}}$$

$$\text{MEAN} = \frac{425}{21} = 20.2$$




ALL OF THE DATA IN THE DISTRIBUTION IS USED IN CALCULATING THE MEAN.

SLIDE 4

The median is the midpoint in a score distribution. It is the point in the distribution of scores where half the number of scores are above and half are below.

MEDIAN

THE MIDPOINT OF THE DISTRIBUTION

8		
9		
9		TEN
13		
13		SCORES
15		
17		ABOVE
17		
17		
20		
22	-	MEDIAN
		
25		
25		TEN
25		
25		SCORES
26		
27		BELOW
27		
27		
28		
30		

SLIDE 5

The median may be useful when scores at either end of the scale will present an unrealistic picture of the average. In this example of 15 scores, the four highest scores are quite different from the rest of the distribution.

Consequently, the mean will be much higher than the median. Though both the mean and the median are averages, there is a considerable difference in the message each reveal.

MEDIAN

SOMETIMES USED WHEN A FEW SCORES WILL UNREALISTICALLY DISTORT THE MEAN.

8		
8		
9		
10		N = 15
10		
11		SUM = 304
11		
11	←	MEAN = 20.3
	MEDIAN = 11	
12		
12		MEDIAN = 11
12		
	←	MEAN = 20.3
	MEAN = 20.3	
42		
46		
50		
52		

SLIDE 6

The mode is the most frequent score in the distribution. It is used when the most frequent number is desired. An example of this use is a school cafeteria manager looking to know which one of six food options is selected by the majority of the children.

MODE

THE MOST FREQUENTLY OCCURRING SCORE IN A DISTRIBUTION OF SCORES.

USED WHEN THE MOST FREQUENT NUMBER OR SIZE OF SOMETHING IS DESIRED.

SLIDE 7

Standard deviation is an index of variability. It will indicate whether scores tend to group close to the mean or spread out in a wider range throughout the distribution. Standard deviation is an important concept in applying data for classroom instructional management. This particular example is used to show a simple method of obtaining standard deviation from a test given to an instructional group of 12 children. The average score (mean) for this class is 10. The scores in this class are ranked from high to low with the central point (mean) indicated with an arrow.

STANDARD DEVIATION

AN INDEX OF VARIABILITY - OR HOW "SPREAD OUT" THE SCORES ARE IN A DISTRIBUTION.

<u>NAME</u>	<u>SCORE</u>	
JOHN JONES	17	
SAM WALLACE	15	
SUE BAKER	13	
ALICE BROWN	12	
JOE MARTIN	12	
JAN DOE	11	
JILL SANGOR	10	← MEAN
BILL BURNS	9	
DAVID DUNN	8	
MARY MILLS	6	
DON ATLEE	5	
SALLY SMITH	<u>2</u>	
SUM	=	120

No. OF PUPILS = 12

$$\text{AVERAGE (MEAN)} = \frac{120}{12} = 10$$

SLIDE 8

The mean (10) is subtracted from each pupil's score which results in a column of difference or deviation scores - the amount each score deviates from the class average. These deviation scores cannot be added in a meaningful way since their sum will be zero. To make these deviation scores useful with respect to the normal or bell-shaped curve, the negative signs must be removed.

STEPS IN DETERMINING STANDARD DEVIATION

<u>NAME</u>	<u>SCORE</u>		<u>AVERAGE (MEAN)</u>	<u>DIFFERENCE (OR DEVIATION) OF SCORES FROM THE MEAN</u>
JOHN JONES	17	-	10	7
SAM WALLACE	15	-	10	5
SUE BAKER	13	-	10	3
ALICE BROWN	12	-	10	2
JOE MARTIN	12	-	10	2
JAN DOE	11	-	10	1
JILL SANGOR	10	-	10	0
BILL BURNS	9	-	10	-1
DAVID DUNN	8	-	10	-2
MARY MILLS	6	-	10	-4
DON ATLEE	5	-	10	-5
SALLY SMITH	2	-	10	-8

MEAN SCORE = 10

MEAN SCORE IS SUBTRACTED FROM EACH PUPIL'S SCORE TO DETERMINE THE DIFFERENCE BETWEEN EACH PUPIL'S SCORE AND THE AVERAGE (MEAN) OF THE GROUP.

SLIDE 9

The negative signs of the deviation in scores are removed by squaring each, which results in the squared deviation column shown on the right. These figures are added, with the sum of 202. The sum of these squared differences (202) is then to be divided by the number of pupils (12) to obtain the average spread of scores around the mean. The results of this calculation is 16.83. However, since the result obtained is based on squared differences in eliminating the negative signs, it is necessary to take the square root of the answer (16.83) to reduce the figure to the more appropriate magnitude of 4.1 -- which is the standard deviation. In deriving the standard deviation, the purpose is to obtain the average of the differences (or deviation) between the scores of a group and the group's mean score. Consequently, in this example of a class of 12 pupils with an average (mean) score of 10, the average difference between the scores and the mean of the group is 4.1 points. If the scores in the group are normally distributed (i.e. follow the pattern of the bell curve), then about 68% of the pupils in this example would fall within a range of approximately 4 points from the mean. In this case, 67% of the pupils (8 out of 12) have scores between 6 and 14. The larger the standard deviation, the greater the spread of scores around the mean. Conversely, the smaller the standard deviation, the closer the scores group around the mean.

<u>NAME</u>	<u>SCORE</u>	<u>MEAN</u>	<u>DIFFERENCE OR DEVIATION</u>	<u>SQUARED DIFFERENCES OR DEVIATIONS</u>
JOHN JONES	17	10	7	49
SAM WALLACE	15	10	5	25
SUE BAKER	13	10	3	9
ALICE BROWN	12	10	2	4
JOE MARTIN	12	10	2	4
JAN DOE	11	10	1	1
JILL SANGOR	→ 10	10	0	0
BILL BURNS	9	10	-1	1
DAVID DUNN	8	10	-2	4
MARY MILLS	6	10	-4	16
DON ATLEE	5	10	-5	25
SALLY SMITH	2	10	-8	64
			<hr/>	<hr/>
			SUM	0
				202

$$\text{STANDARD DEVIATION} = \sqrt{\frac{202}{12}}$$

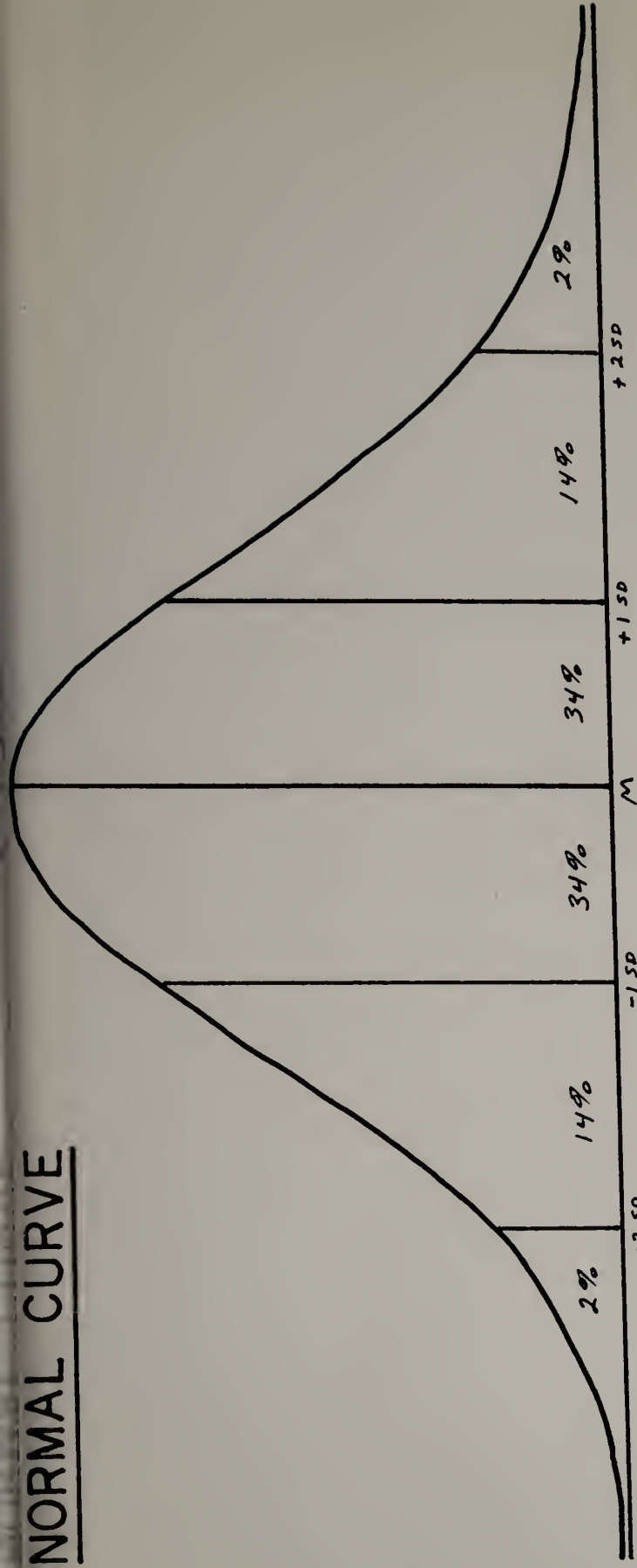
$$= \sqrt{16.83}$$

$$= 4.1$$

SLIDE 10

The normal curve (or the Bell-shaped curve) is an important part of norm-referenced test data interpretation. The normal curve is composed of a central point (mean) in the total distribution of scores. The remaining vertical lines indicate positions of the standard deviations away from the mean. If a distribution of scores follows the pattern of normal distribution, approximately 68% of the scores will fall within one standard deviation in both directions from the mean. Different types of score scales are shown which reveal both unequal and equal scaling properties. Percentiles and grade equivalent scores divide the normal distribution into unequal segments, whereas the differences between the standard scores are the same across an entire distribution. The feature of standard scores dividing the normal curve into equal segments make them useful in determining pupil growth characteristics. It also makes them more amenable to statistical analysis since they can be dealt with mathematically. Grade equivalent scores and percentiles cannot be dealt with mathematically - for example, they should not be averaged.

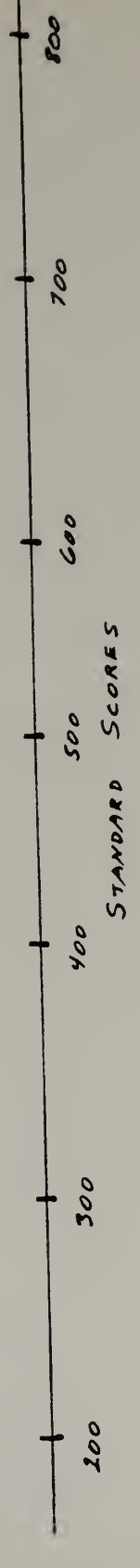
NORMAL CURVE



PERCENTILE	5	10	20	30	40	50	60	70	80	90	95	99
------------	---	----	----	----	----	----	----	----	----	----	----	----

GRADE EQUIV.	2.4	3.3	3.6	4.6	5.0	5.5	6.0	6.7	7.3	8.1	9.1	10.6	11.9
--------------	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	------	------

DIFFERENCE	.9	.3	1.0	.4	.5	.5	.7	.6	.8	1.0	1.5	1.5	1.3
------------	----	----	-----	----	----	----	----	----	----	-----	-----	-----	-----



SLIDE 11

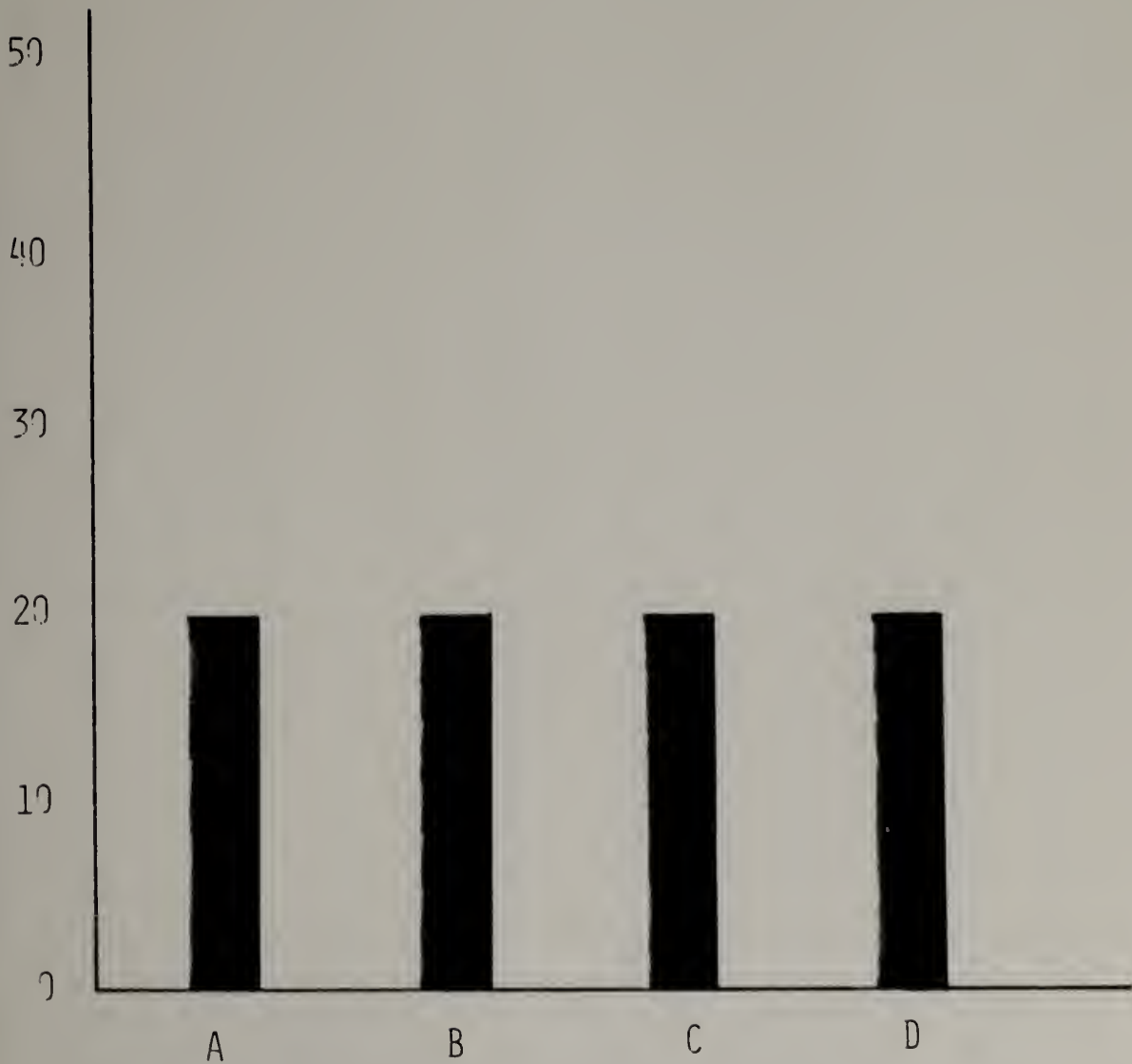
The shape of a distribution, in addition to the central point (mean) and the spread of scores (standard deviation) around a central position, can have instructional significance. There is a tendency for the policy making bodies (school administrators, teachers and parents) in a school district to focus on the average score as the index of academic success or failure. It is often necessary to look beyond the average for indices of group performance characteristics to make more effective instructional decisions. Standard deviation is an additional index of the degree of score variability. Another index is skewness, or the shape of a score distribution. The following transparencies show how skewness can be used as a technique of gathering additional information for instructional management.

This graphic display of average (mean) scores for four classes at Westside School reveals similar means. Given the fact that each class has 30 pupils and each has the same average (mean) score - are all four classes alike? Is there any information that can tell us that in fact the four classes are different? Or are they alike?

READING TEST RESULTS

WESTSIDE SCHOOL

GRADE 5



SLIDES 12-16

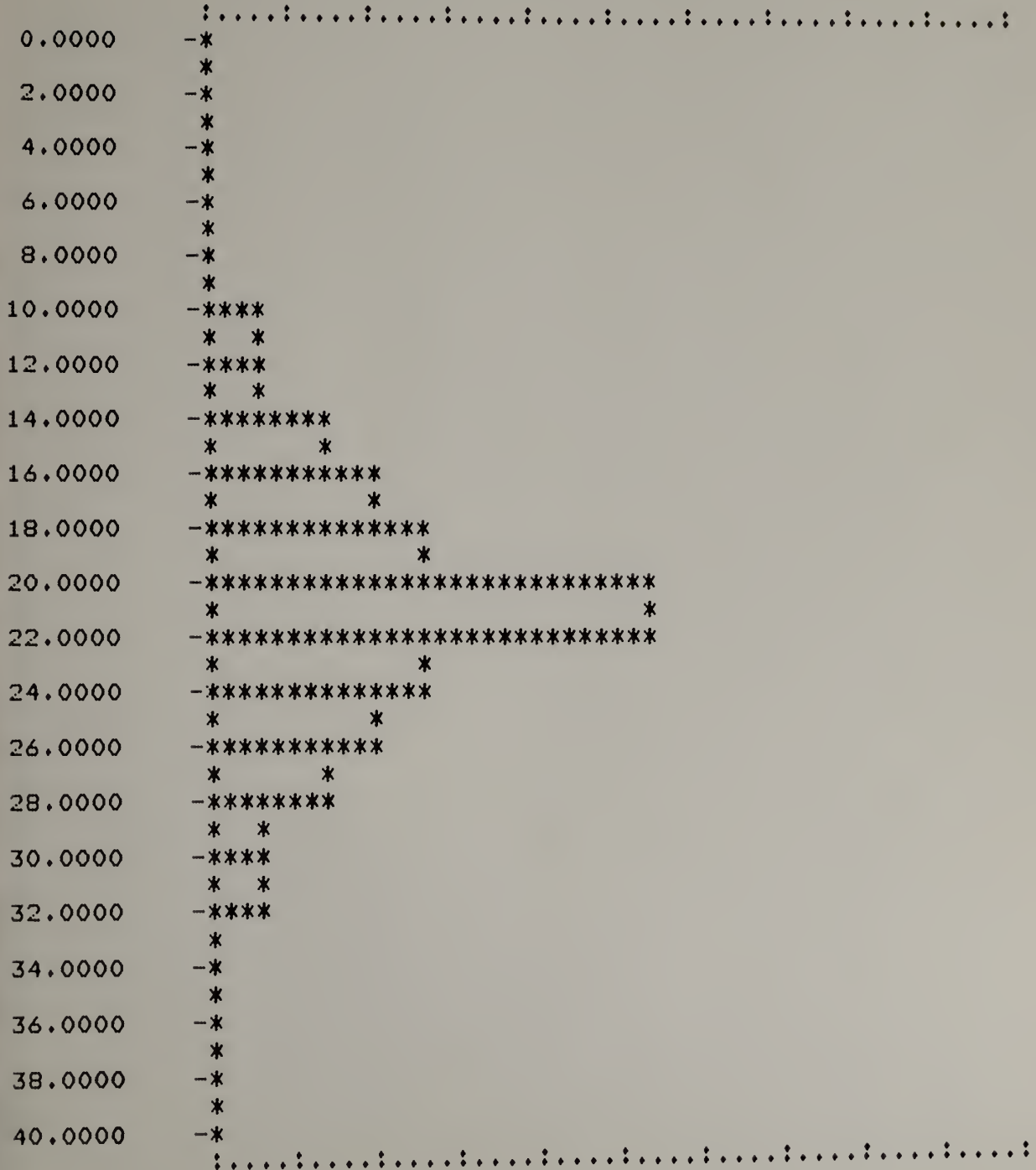
Data from the Westside classes are shown in the following transparencies. Though the mean scores and the number of children tested are the same in every case, the standard deviations reflect differences. In addition, each score distribution has an entirely different shape. The emphasis here is placed on the fallability of the average (mean) score as a sole index of group performance. Averages tend to be insensitive to patterns of data within the distribution, which can reveal significant information for classroom instructional management.

<u>NORMAL</u>	<u>SKewed</u>	<u>COMPRESSED</u>	<u>BI-MODAL</u>
10	13	19	10
12	13	19	11
15	15	19	12
15	15	19	13
16	15	19	13
16	16	19	14
16	16	19	14
18	16	19	14
18	17	19	14
18	18	19	14
18	18	20	15
20	19	20	15
20	19	20	16
20	20	20	17
20	21	20	18
20	21	20	20
20	21	20	24
20	22	20	25
20	22	20	25
22	22	20	26
22	22	21	26
22	22	21	26
22	24	21	26
24	24	21	26
24	24	21	26
24	25	21	27
27	25	21	27
27	25	21	28
28	25	21	29
30	25	21	30
N = 30	N = 30	N = 30	N = 30
MEAN = 20	MEAN = 20	MEAN = 20	MEAN = 20
SD = 4.6	SD = 3.9	SD = .8	SD = 6.5

NORMAL
DISTRIBUTION
N = 30
MEAN = 20.0
SD = 4.6
SKEWNESS = 0

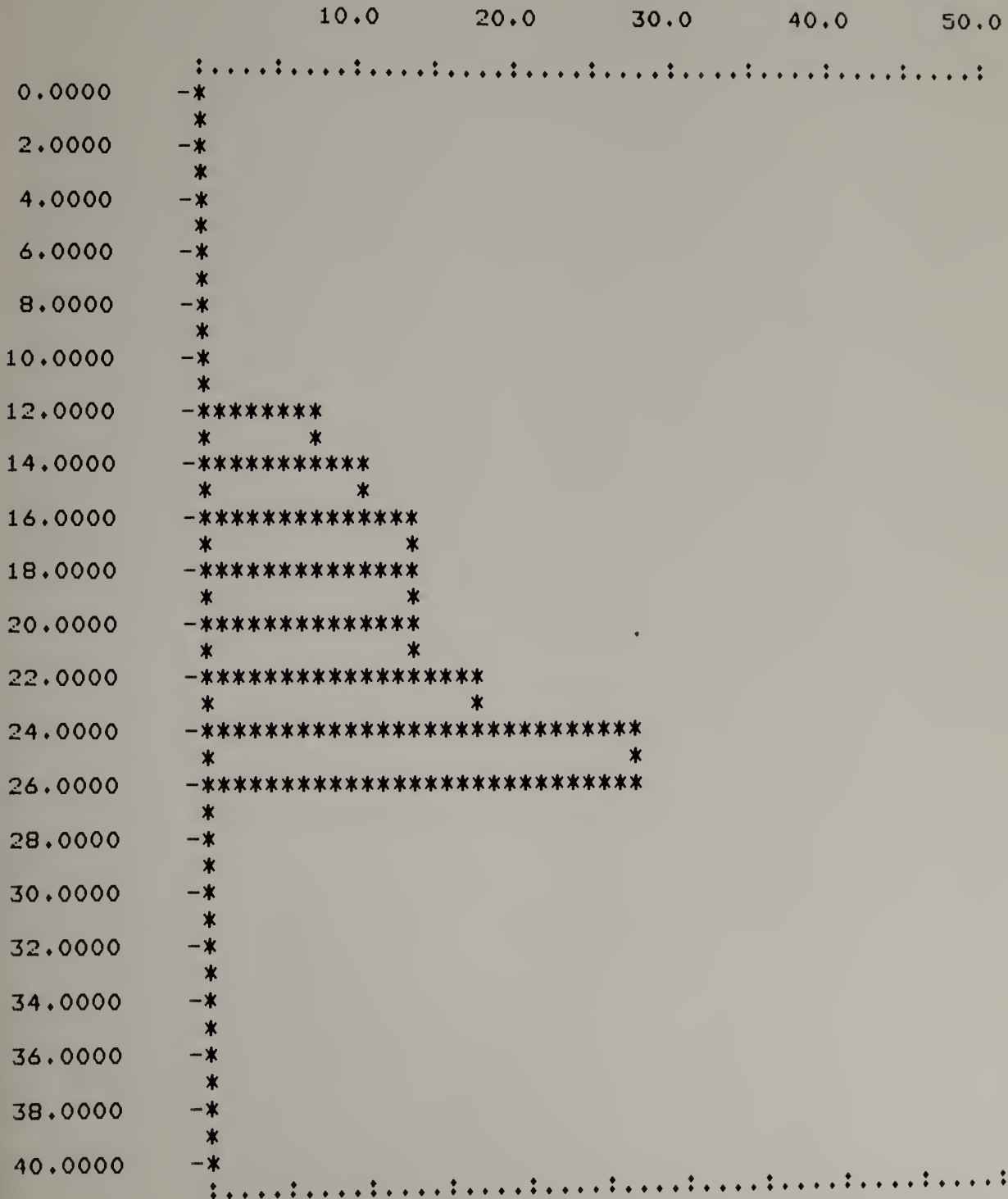
RELATIVE FREQUENCY

10.0 20.0 30.0 40.0 50.0



SKEWED
 DISTRIBUTION
 N = 30
 MEAN = 20.0
 SD = 3.9
 SKEWNESS = -2.635

R E L A T I V E F R E Q U E N C Y



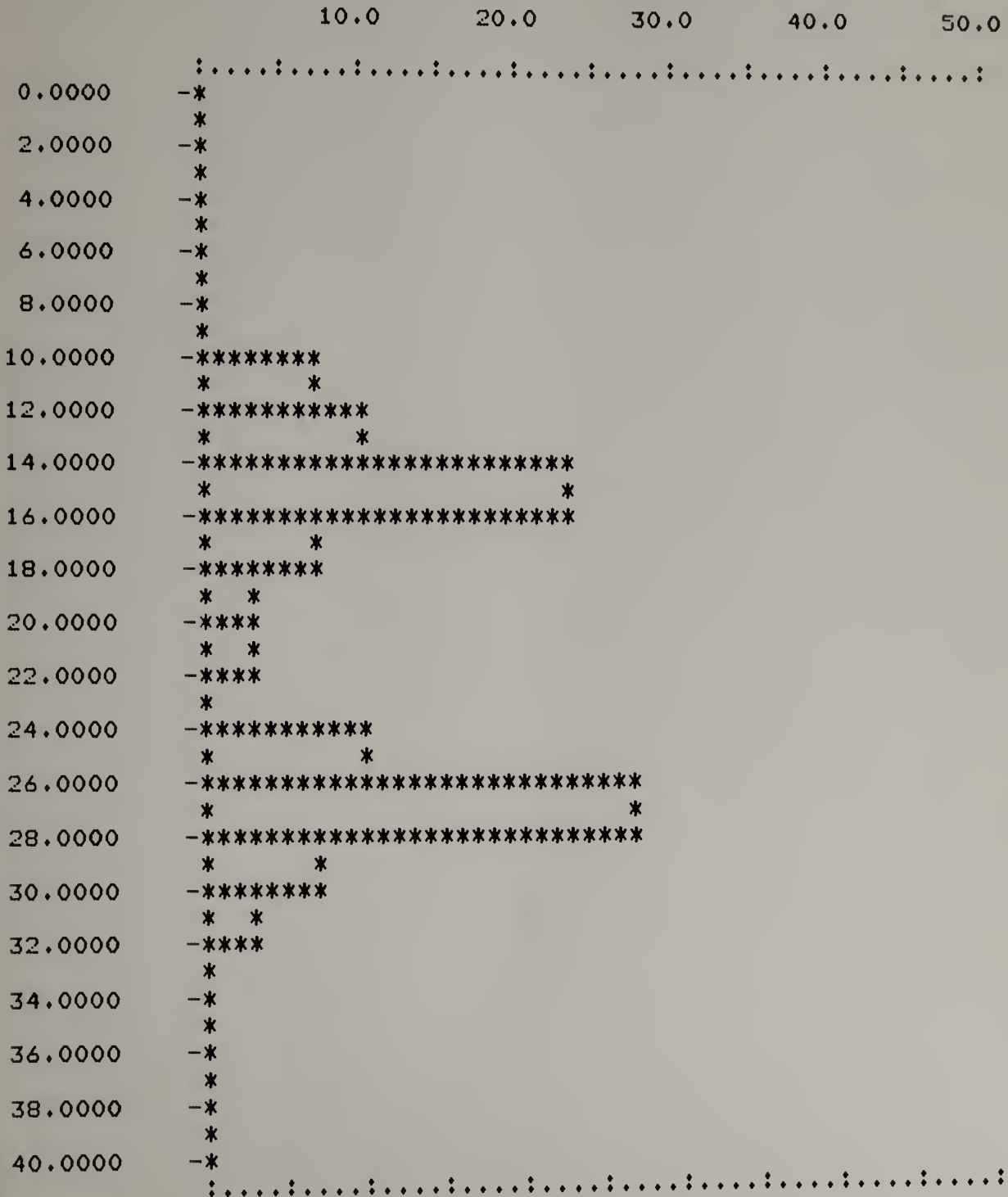
COMPRESSED
 DISTRIBUTION
 N = 30
 MEAN = 20.0
 SD = .830
 SKEWNESS = 0

RELATIVE FREQUENCY



BI-MODAL
 DISTRIBUTION
 N = 30
 MEAN = 20.0
 SD = 6.5
 SKEWNESS = 0

RELATIVE FREQUENCY



SLIDE 17

A computer isn't needed to generate frequency distributions or graphic profiles of group performance. A classroom teacher can take the scores for a group, list them vertically from low to high, and place a symbol for each pupil achieving the respective scores. In this way, a distribution of scores can be graphically portrayed for grouping or other instructional purposes.

Additional information can be gained from a teacher created frequency distribution by recording a unique symbol for each child such as the child's initials rather than the x's shown in this example. In this way, if the frequency distribution reveals any grouping pattern such as the one seen in this example then the teacher can easily identify pupils for instructional grouping purposes.

TEACHER MADE FREQUENCY DISTRIBUTION

	<u>FREQUENCY</u>
10	X
11	X
12	X
13	X X
14	X X X X X
15	X X
16	X
17	X
18	X
19	
20	X
21	
22	
23	
24	X
25	X X
26	X X X X X X
27	X X
28	X
29	X
30	X

SLIDE 18

The correlation coefficient reflects the degree of relationship between two or more variables. The total range for a correlation coefficient is from -1.00 (indicating a perfect negative relationship) to +1.00 (indicating a perfect positive relationship) with a .00 correlation coefficient reflecting no significant relationship between the variables. The correlation coefficient is usually used to give evidence of the reliability and/or validity of a test. Reliability may be determined by correlating two administrations of a test to the same group of pupils within short intervals of time with no intervening instruction. If the resulting correlation coefficient is positive and high, i.e. .95, this indicates the response patterns of pupils in two testing situations is consistent. Validity may be determined by correlating results of a locally developed test with scores of the same pupils on a nationally recognized test measuring the same skills. A resulting positive and high correlation coefficient would support the idea that children are performing in a similar fashion on both instruments. The correlation coefficient is frequently misinterpreted as showing that one variable may be causing the relationship with another variable. Simply because one variable may be highly correlated with another is no indication that one variable may be the cause of this high correlation. For example, there may be a high positive correlation between shoe size and scholastic aptitude but that does not mean one needs big feet in order to be successful in school.

CORRELATION

THE COEFFICIENT OF CORRELATION IS A NUMBER INDICATING THE DEGREE OF RELATIONSHIP BETWEEN TWO VARIABLES. THE RANGE FOR THE CORRELATION COEFFICIENT IS FROM -1.00 TO +1.00.

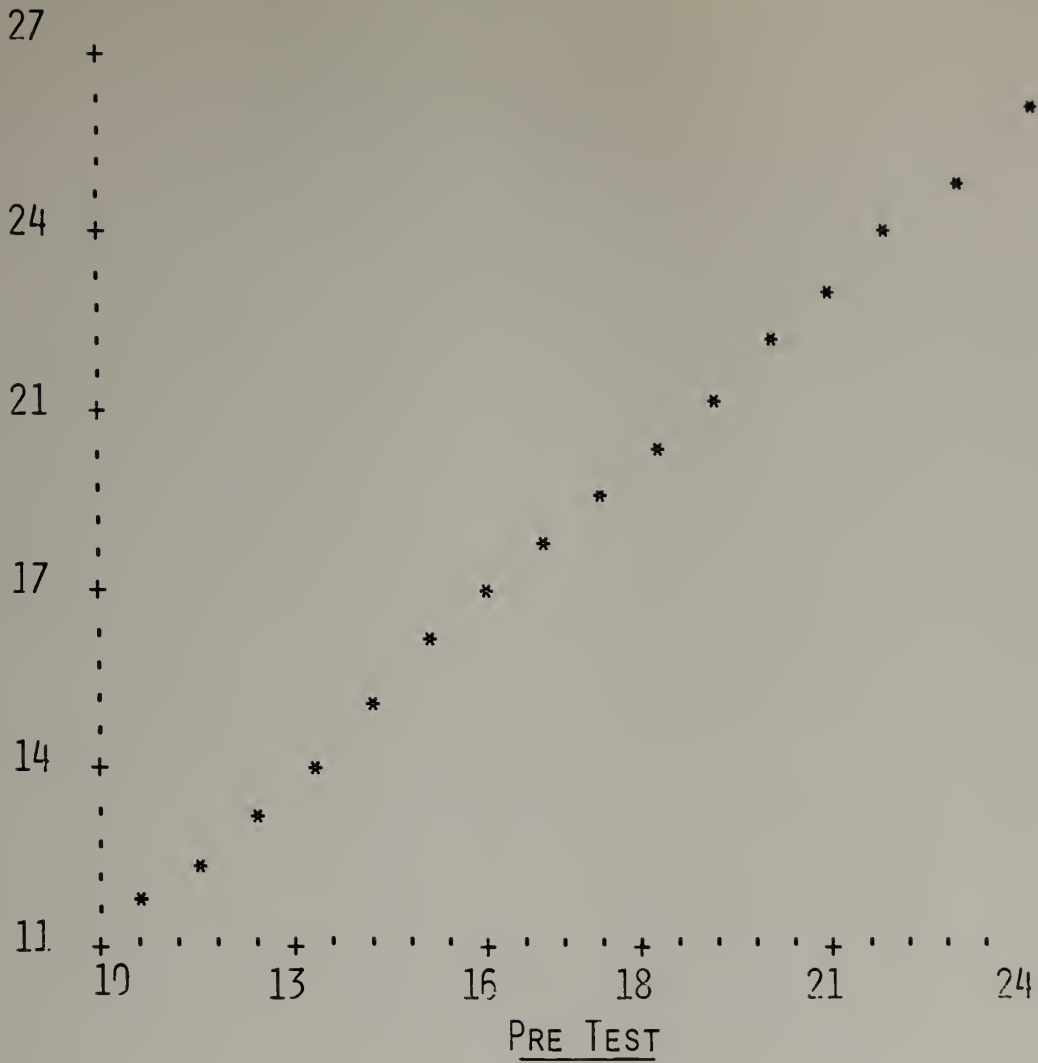
-1.00	PERFECT NEGATIVE RELATIONSHIP
- .99	
- .98	
.	
.	
.00	NO RELATIONSHIP
.	
.	
+ .98	
+ .99	
+1.00	PERFECT POSITIVE RELATIONSHIP

- CORRELATION IS NOT EVIDENCE OF CAUSATION.
- CORRELATION USED IN DETERMINING TEST RELIABILITY AND VALIDITY.

SLIDE 19

This transparency indicates a situation where there is a perfect positive correlation. Two variables, in this case the pretest and posttest, are perfectly related. The correlation coefficient between the two is +1.00. Pupils who receive a low score on the pretest are the same pupils who receive low scores on the posttest. Pupils who receive high scores on the pretest are also the same pupils who receive high scores on the posttest. This degree of relationship is consistent across the score scale for both pre and post test results.

POST TEST



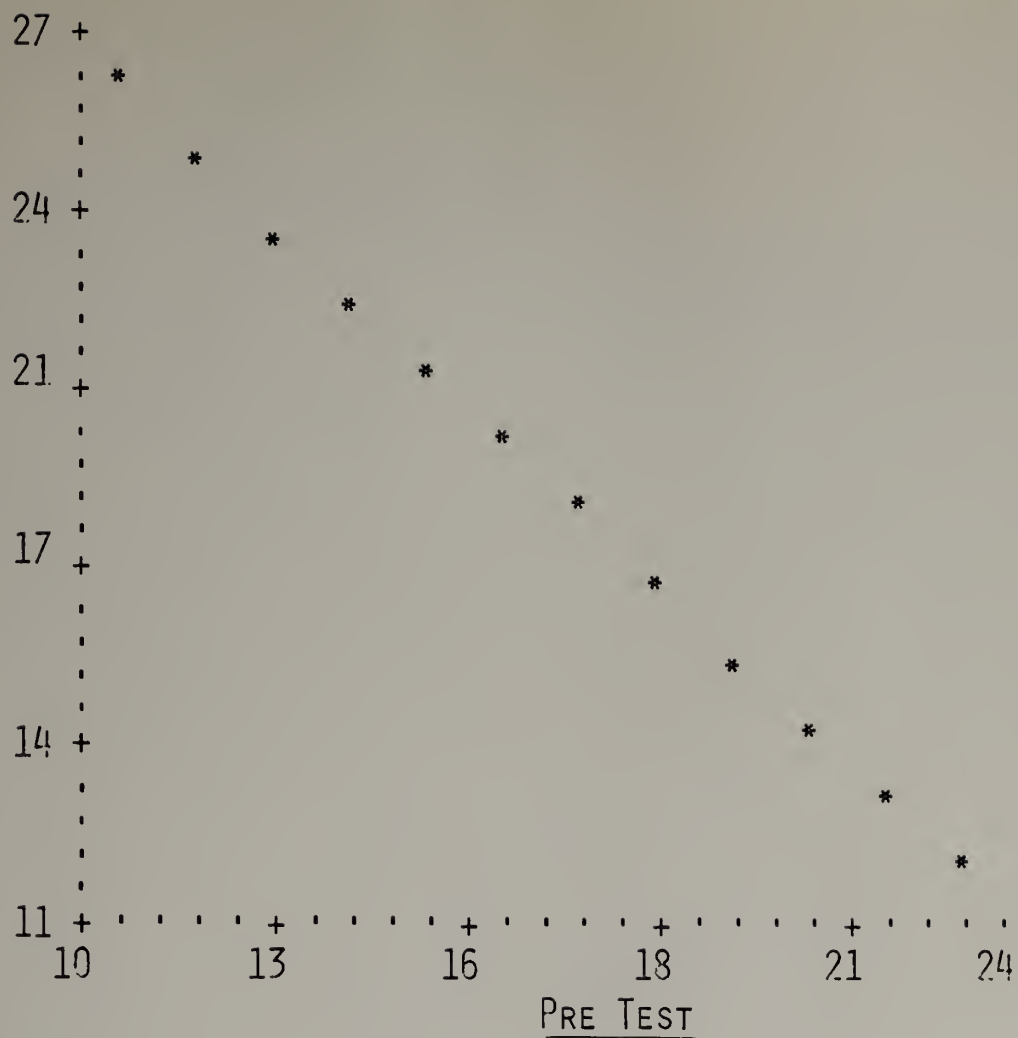
<u>PRE TEST</u>	<u>POST TEST</u>
10	12
11	13
12	14
13	15
14	16
15	17
16	18
17	19
18	20
19	21
20	22
21	23
22	24
23	25
24	26

CORRELATION = +1.000

SLIDE 20

This transparency shows a perfect negative correlation. In this case, pupils who have low pretest scores have high posttest scores, and pupils who have high pretest scores have low posttest scores. This particular situation is consistent throughout the score scale and results in a perfect inverse relationship between the two variables.

POST TEST



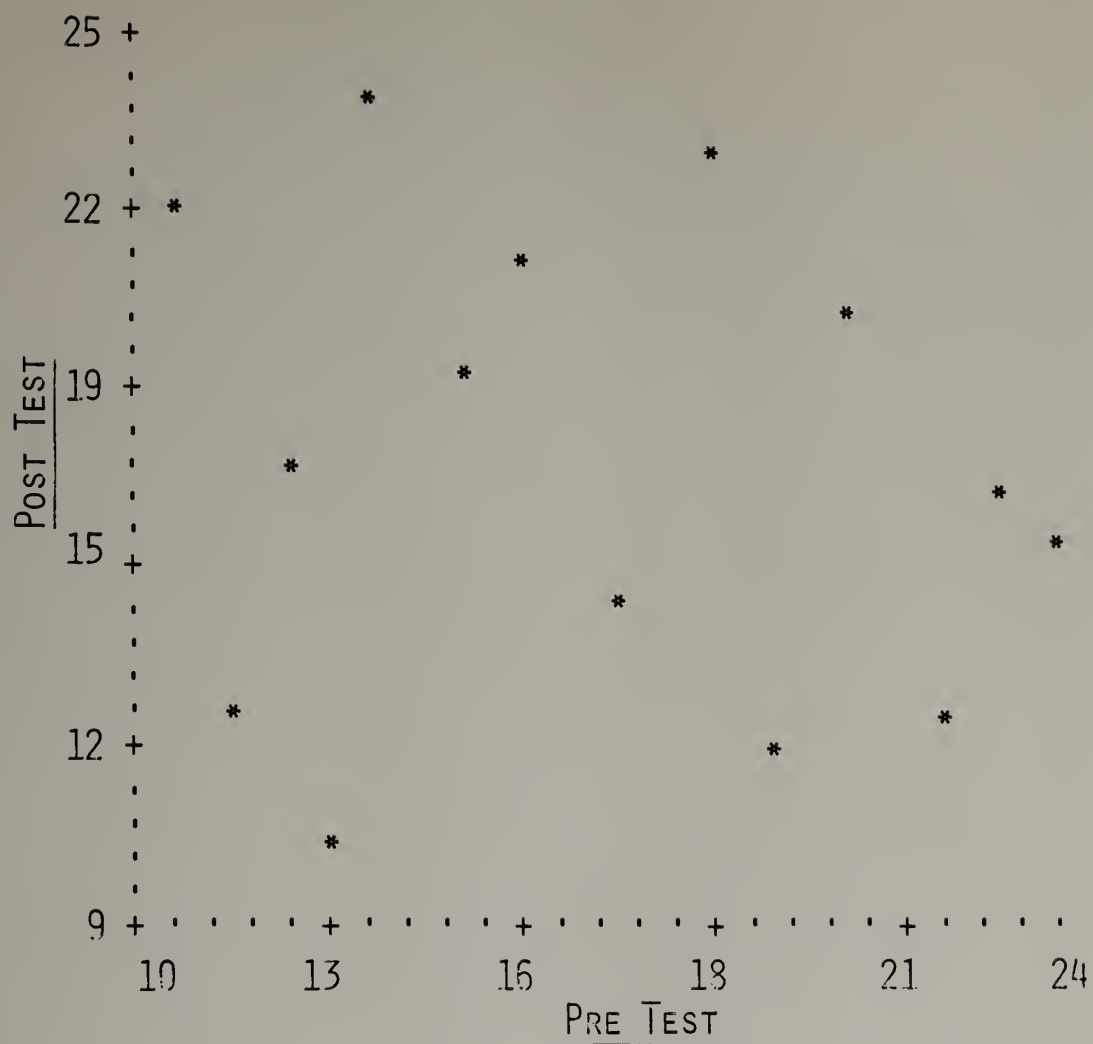
PRE TEST POST TEST

10	22
11	25
12	24
13	23
14	22
15	21
16	20
17	19
18	18
19	17
20	16
21	15
22	14
23	13
24	12

CORRELATION = -1.000

SLIDE 21

This transparency indicates virtually no correlation between two variables.
The students with low pretest scores may or may not have low posttest scores.
Pupils with high pretest scores may or may not have high posttest scores.
There is no consistent pattern of performance from pretesting to posttesting
resulting in no relationship between the testing situations.



<u>PRE TEST</u>	<u>POST TEST</u>
-----------------	------------------

10	22
----	----

11	13
----	----

12	17
----	----

13	10
----	----

14	24
----	----

15	19
----	----

16	21
----	----

17	14
----	----

18	23
----	----

19	11
----	----

20	20
----	----

21	18
----	----

22	12
----	----

23	16
----	----

24	15
----	----

CORRELATION = $-.17$

SLIDE 22

Reliability indicates how consistently a test measures what it is supposed to measure. If pupils are tested at a particular time on a particular test, and then are tested again without any intervening instructions, will they get roughly the same scores they had the first time? This is the basic question of reliability. It measures consistency of response patterns from one testing situation to another without any influence of intervening instruction.

Reliability can be obtained in several ways. One way is by correlating scores on one form of a test with scores from the same pupils on another form of the test. Another method of determining reliability is to give one test to a group of pupils, and correlate the results of one half with the results of the other half of the test. A third procedure for establishing reliability is to administer the same test twice to the same group of pupils and determine the degree of relationship between the two sets of scores. Reliability, therefore, is determined by correlating data obtained in one testing situation with data obtained through another testing situation whether it be an alternative form, a different half of the same test, or the same test administered twice. Examples of reliability coefficients for a particular standardized norm-referenced achievement test are given and discussed with respect to their derivation and applicability toward greater understanding of the use of test data. Users of the test are cautioned to look for indices of reliability to give them further information on how much confidence they should place on both individual and group test results.

RELIABILITY

HOW CONSISTENTLY A TEST MEASURES WHAT IT IS SUPPOSED TO MEASURE.

DETERMINED BY CORRELATING:

1. SCORES ON ALTERNATE AND EQUIVALENT FORMS.
2. SCORES ON TWO HALVES OF THE SAME TEST.
3. SCORES ON THE SAME TEST ADMINISTERED TWICE.

<u>SUBTEST</u>	<u>RELIABILITY COEFFICIENT</u>
READING VOCABULARY	.92
READING COMPREHENSION	.91
READING TOTAL	.95
SPELLING	.87
LANGUAGE MECHANICS	.74
LANGUAGE EXPRESSION	.85
TOTAL LANGUAGE	.92

SLIDE 23

The purpose of validity is to determine how well a test measures what it is supposed to measure. There are several types of validity - three which are mentioned here. Content validity is an important consideration when dealing with norm-referenced standardized tests, particularly if there is pressure placed on using these tests to evaluate the overall curricular in a school district. Content validity involves a subjective process where the other two types of validity make use of data in the form of correlation coefficients to support their significance.

VALIDITY

HOW WELL A TEST MEASURES WHAT IT IS SUPPOSED TO MEASURE.

CONTENT VALIDITY

HOW WELL THE ITEMS OF A TEST REPRESENT THE PARTICULAR SEQUENCE OR TYPES OF SKILLS BEING TESTED. A SUBJECTIVE APPROACH.

CRITERION-RELATED VALIDITY

HOW WELL A TEST TO BE VALIDATED CORRELATES WITH AN ESTABLISHED INDEPENDENT CRITERION WHICH IT WAS DESIGNED TO MEASURE. COLLEGE BOARDS HAVE CRITERION-RELATED VALIDITY IN THAT THEIR SCORES TEND TO BE INDICATIVE OF LATER SUCCESS IN COLLEGE.

CONSTRUCT VALIDITY

HOW WELL THE RESULTS OF A TEST RELATE TO OTHER MEASURES RESEARCH HAS SHOWN TO HAVE THE SAME THEORETICAL FRAMEWORK. A PERSON SCORING HIGH IN ANXIETY ON ONE MEASURE WOULD BE EXPECTED TO SCORE HIGH IN ANXIETY ON OTHER MEASURES OF TRAITS THAT RESEARCH HAS SHOWN TO BE ASSOCIATED WITH ANXIETY.

SLIDE 24

The issues of reliability and validity with criterion-referenced tests present a different set of problems. These definitions are extracted from Popham, W. J., Educational Evaluation, Englewood Cliffs, New Jersey; Prentice-Hall, 1975, Chapter 7. Emphasis is placed on the non-statistical approach used in defining these concepts in the criterion-referenced domain.

CRITERION REFERENCED MEASUREMENT

RELIABILITY

CONSISTENCY OF INSTRUCTIONAL ASSIGNMENTS MADE ON THE BASIS OF CRITERION REFERENCED MEASUREMENT RESULTS.

CONSISTENCY OF ITEM RESPONSE PATTERNS ON TWO DIFFERENT TESTINGS - OR THE PERCENTAGE OF PUPILS' SCORES THAT MAY VARY ON TWO DIFFERENT TESTINGS.

DETERMINED BY HOW WELL THE ITEMS RELATE TO A GIVEN DOMAIN - SHOULD BE A REQUIRED PROCESS IN DEVELOPING CRITERION REFERENCED MEASUREMENT.

VALIDITY

DESCRIPTIVE - DO THE TEST ITEMS ADEQUATELY DESCRIBE THE DOMAIN TO BE EXAMINED?

FUNCTIONAL - DOES THE TEST ACTUALLY ACCOMPLISH WHAT IT IS SUPPOSED TO ACCOMPLISH - IS IT SENSITIVE TO INSTRUCTION?

DOMAIN-SELECTION VALIDITY - HOW ACCURATE IS THE DOMAIN SELECTED? HOW GENERALIZABLE WILL THE RESULTS BE AS AN INDICATOR OF LEARNER STATUS WITH RESPECT TO A MORE GENERAL DIMENSION?

APPLICATION

IMPROVING INSTRUCTION THROUGH UNDERSTANDING
AND APPLYING DATA

SLIDE 1

This section of the staff development program is concerned with applying test data to improving instruction. The approach used in this section lists and defines different types of scores obtained through testing, shows the advantages and disadvantages of each, and shows how they can be used with other relevant data to improve instructional practices. These scores will be amplified with examples from existing norm-referenced standardized tests. However, emphasis will be placed on how the scores can also be applied to teacher-made tests.

TYPES OF SCORES

- RAW SCORES

- GRADE EQUIVALENT SCORES

- STANDARD SCORES

- PERCENTILES
 - NATIONAL
 - LOCAL

- EXPECTANCY SCORES

SLIDE 2

Raw scores represent the total number of items each child answered correctly on a given test. Raw scores have little meaning used alone. More information about the distribution and range of scores is needed to make meaningful use of raw scores.

RAW SCORES

TOTAL NUMBER OF CORRECT ITEMS

VARIES FOR EACH SUBTEST DEPENDING ON THE
NUMBER OF ITEMS

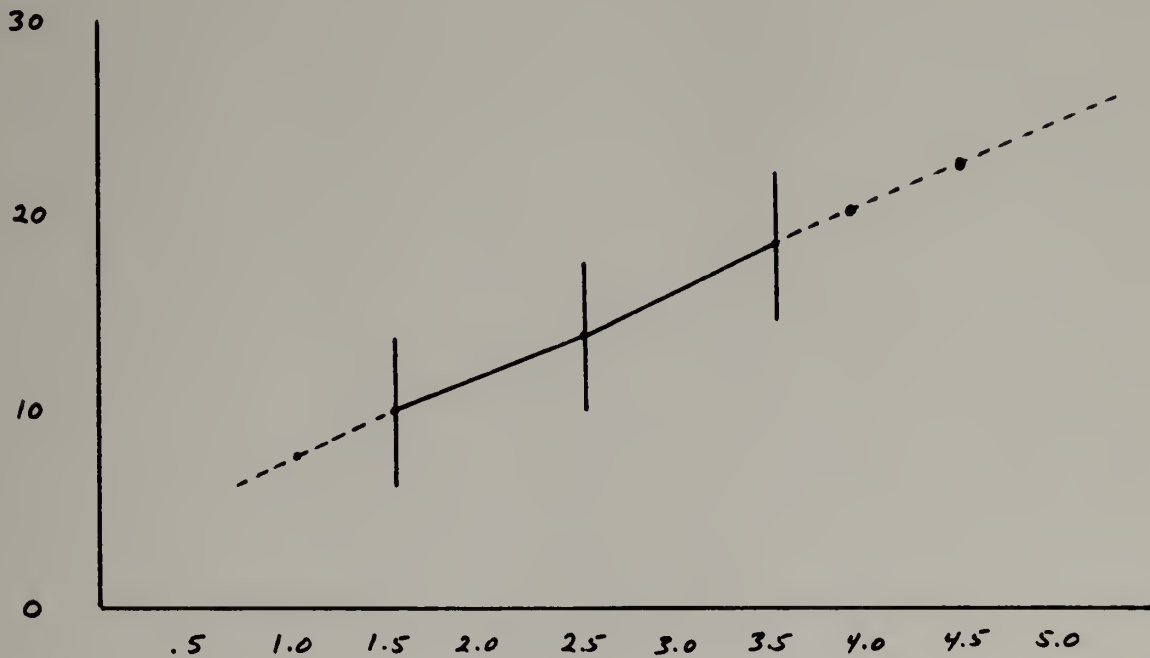
DO NOT RELATE, BY THEMSELVES, TO A REFERENCE
GROUP.

SLIDE 3

The grade equivalent score is essentially a placement in terms of school grade, in years and months, which the raw score is an actual or estimated average. One thing frequently not understood about grade equivalent scores is that a given grade equivalent score is not necessarily the average score obtained by all students at a particular grade level. Most publishers of norm-referenced tests develop norms for a test based on giving the test at a particular time of year. The vertical lines on this graph might represent the positions of the average scores for each respective grade placement (1.5, 2.5 and 3.5) at the time the test was administered. Grade equivalent scores between actual tested grade level placement is obtained through interpolation. Grade equivalent scores for points in the distribution beyond the grade levels actually tested is obtained through extrapolation. The assumption is made that learning results in a steady linear pattern throughout each and every year, including summer.

GRADE EQUIVALENT SCORES

THE GRADE PLACEMENT IN SCHOOL YEARS AND MONTHS FOR WHICH THE OBTAINED RAW SCORE IS THE ACTUAL OR ESTIMATED AVERAGE (AVERAGE IS EITHER MEAN OR MEDIAN).



<u>RAW SCORE</u>	<u>G.E.</u>	<u>HOW OBTAINED</u>
8	1.0	EXTRAPOLATION
9	1.3	EXTRAPOLATION
10	1.5	ACTUAL DATA
11	1.8	INTERPOLATION
12	2.2	INTERPOLATION
18	3.5	ACTUAL DATA
19	3.9	EXTRAPOLATION
21	4.5	EXTRAPOLATION

SLIDE 4

Grade equivalent scores are often misused. A major problem with these scores is their apparent index of grade placement. Problems in using these scores are described below:

1. A 6th grade child whose grade equivalent score on a test is 8.5 is often considered capable of doing the same quality of school work as children who are in the 5th month of the 8th grade. The grade equivalent score of 8.5 for this child reflects an above average position on this test's scale, not necessarily that the child should be in the middle of the 8th grade.

2. Grade equivalent scores and their score scale assume growth is consistent throughout the school year; a situation which may not reflect reality. Studies have shown growth patterns to reflect considerable variability throughout the year, and the patterns will differ for different children.

3. Since grade equivalent scores are based on a scale which breaks the scores down into unequal segments, it is not possible to legitimately average these data. Consequently, group performance or comparing groups using grade equivalent score averages may reveal a completely unrealistic picture of group progress.

4. There is a large within grade variability in the grade equivalent score scale which can result in an average performer obtaining a score a year or two above his grade placement. A slight fluctuation in the number of correct items on a particular subtest can result in a large grade equivalent score gain.

5. Grade equivalent scores are frequently misinterpreted as a desirable standard. A child with a grade equivalent score equal to his actual grade placement may be considered to be working at a proper grade level and not require any additional help. It is possible that a child who is tested at

SLIDE 4 (Continued)

grade level could be performing in school above grade level if given additional instructional support.

6. Grade equivalent scores are frequently used to compare children's performance on different tests. A child with a particular grade equivalent score on one subtest and a similar grade equivalent score on another may be considered as working at equal proficiency on both subtests when in fact one subtest may be scaled differently than the other. Grade level for one subtest may involve a relatively lower level of proficiency than grade level on another subtest.

PROBLEMS WITH GRADE EQUIVALENT SCORES

1. GRADE EQUIVALENT SCORES MAY NOT ACCURATELY REFLECT A PROPER MESSAGE.
2. THE SCORES ASSUME CHILDREN LEARN AT THE SAME RATE THROUGHOUT THE YEAR - INCLUDING SUMMER.
3. THE SCORES ARE BASED ON A SCALE COMPOSED OF UNEQUAL SEGMENTS WHICH MAKES GROUP AVERAGES AND COMPARISONS UNREALISTIC.
4. LARGE VARIABILITY WITHIN THE GRADE EQUIVALENT SCORE SCALE CAN RESULT IN AN "AVERAGE" TEST PERFORMER APPEARING TO BE A YEAR OR TWO ADVANCED.
5. GRADE EQUIVALENT SCORES ARE FREQUENTLY MISINTERPRETED AS BEING A "STANDARD" WHICH COULD RESULT IN EITHER UNREALISTIC GOALS OR COMPLACENCY WHEN, IN FACT, PERFORMANCE COULD BE HIGHER.
6. GRADE EQUIVALENT SCORES ARE OFTEN USED TO COMPARE CHILDREN'S PERFORMANCE ON DIFFERENT TESTS.

SLIDE 5

This transparency shows examples of popular misuses of grade equivalent scores. In the first case, Suzy in grade 3 has a reading grade equivalent score of 5.5. The question asked is does she read as well as a fifth grade fifth month child? The answer is that she does well on the reading skills measured by this test. Her performance is similar to the average 5th grade 5th month child on the third grade material in this test. This assumes the norming sample for this test used children who were in the 5th month of the 5th grade. Another problem is in the case of Sam who has a grade equivalent reading score of 5.7 and a grade equivalent math score of 4.5. The question may be asked - is Sam better at reading than at math? Reading grade equivalent scales are usually higher than math grade equivalent scales so Sam could, in fact, be equal in both with respect to his grade in school. One of the difficulties in assessing growth or group progress in terms of program evaluation with grade equivalent scores is the fact that pupils will appear to "grow" at different rates depending upon their particular position in the distribution of scores. Children who may pre test at very low ranges may tend to grow very rapidly in terms of post test scores. This could be a result of regression to the mean and may not necessarily be evidence of major academic achievement.

SUZY IS IN GRADE 3
READING G.E. OF 5.5

DOES SHE READ AS WELL AS A 5-YEAR 5-MONTH CHILD?

SHE DOES WELL ON THE READING SKILLS MEASURED BY THIS TEST -
AS WELL AS THE AVERAGE 5.5 GRADER DOES ON 3RD GRADE
MATERIAL, ASSUMING THE NORMING SAMPLE INCLUDED CHILDREN IN
THE 5TH MONTH OF THE 5TH GRADE.

SAM - G.E. READING = 5.7
G.E. MATH = 4.5

IS SAM BETTER AT READING THAN MATH?

READING G.E. SCALE IS USUALLY HIGHER THAN MATH G.E. SCALE -
SO SAM COULD BE EQUAL IN BOTH WITH RESPECT TO HIS GRADE IN
SCHOOL.

PUPILS WILL APPEAR TO "GROW" AT DIFFERENT RATES DEPENDING
ON THEIR POSITION IN THE DISTRIBUTION.

SLIDE 6

At the beginning of the third grade, Tom, Dick and Harry each tested at different percentile ranks. Tom is at the 15th percentile rank, Dick is at the 50th and Harry is at the 85th. Their respective GE scores are 2.0, 3.1 and 5.0. In grade 4 they maintained the same percentile rank but have quite different grade equivalent scores. At grade 5 they again maintain their percentile ranks but Tom over the two-year time span has grown 1.3 years. Dick has grown the 2.0 years since he was in the average percentile and has attained the expected growth pattern. Harry, however, has grown 3.4 years in terms of grade equivalent score gain. Since each is performing at different levels of the score distribution and has maintained the same relative position, their grade equivalent score gain has demonstrated different growth rates for each individual.

Data taken from:

Wick, J. W., Educational Measurement: Where are we going and how will we know when we get there., Columbus, Ohio: Charles E. Merrill Pub. Co., 1973

<u>PUPIL</u>	PERCENTILE RANK AT EACH GRADE LEVEL			
	<u>GRADE 3</u>	<u>GRADE 4</u>	<u>GRADE 5</u>	
TOM	15	2.0	2.7	3.3
DICK	50	3.1	4.1	5.1
HARRY	85	5.0	6.6	8.4

- SCORES ARE LESS STABLE AT THE EXTREMES OF THE DISTRIBUTION.
- STANDARD DEVIATION OF G.E. SCORES INCREASE AS THE GRADES INCREASE.

<u>PUPIL</u>	<u>TWO-YEAR "GROWTH"</u>
TOM	1.3
DICK	2.0
HARRY	3.4

Wick, J. W., Educational Measurement: Where are we going and how will we know when we get there., Columbus, Ohio: Charles E. Merrill Pub. Co., 1973.

SLIDE 7

A standard score is one based on an equal interval scale using a designated mean and standard deviation. This transparency defines how the standard score is computationally derived. Given a group with an average score of 65 and a standard deviation of 15, Johnny has a score of 77. His score in standard deviation units is positive .8. Johnny is .8 standard deviation units away from the class average (mean). This is derived by subtracting the mean of the class from his score and dividing the result (12) by the standard deviation (15). Since it is awkward to deal in decimals, the decimal is eliminated by multiplying .8 by 10. A constant (such as 500) is added to the result in order to eliminate the possibility of negative numbers with low scoring pupils. Consequently, standard scores can indicate the child's relative position in a distribution away from the average, providing the mean and standard deviation is known.

STANDARD SCORE

A SCORE BASED ON A SCALE WITH EQUAL INTERVALS - A SCALE WITH A DESIGNATED MEAN AND STANDARD DEVIATION.

EXAMPLE

CLASS AVERAGE (MEAN) OF A GROUP OF SCORES	65
CLASS STANDARD DEVIATION	15
JOHNNY'S SCORE	77
JOHNNY'S SCORE IN STANDARD DEVIATION UNITS	+.8
JOHNNY IS 12 POINTS ABOVE THE MEAN; THEREFORE, $12 - 15 = .8$	
MULTIPLY JOHNNY'S SCORE IN STANDARD DEVIATION UNITS BY 10 TO ELIMINATE DECIMALS	8
ADD A CONSTANT SUCH AS 500	508
JOHNNY'S STANDARD SCORE	508

SLIDE 8

This transparency shows a practical application of standard scores used to equate a child's performance on two tests. Though Suzy received different raw scores on different tests, in terms of how the groups performed she is in the same relative position from the class average on both tests.

	<u>READING TEST</u>	<u>MATH TEST</u>
CLASS AVERAGE (MEAN)	30	80
CLASS STANDARD DEVIATION	5	20
SUZIE'S SCORE	33	92
DIFFERENCE BETWEEN SUZIE'S SCORES AND THE CLASS AVERAGE	+3	+12
SUZIE'S SCORES IN SD UNITS	$3 \div 5 = .6$	$12 \div 20 = .6$
MULTIPLIED BY 10	6	6
ADD CONSTANT 50	<u>56</u>	<u>56</u>

SLIDE 9

The percentile is the most common index used in norm-referenced standardized tests. It depicts the position of the child in relation to a comparison group. The national percentile compares a child with respect to a national sample. The local percentile depicts the position of the child with respect to his or her peers on a local distribution of scores. The important thing to remember about percentiles is they are not scores themselves, but rather represent the ranking of a child with respect to a defined group. In the case of national percentile this defined group is the national sample. It is important to understand national percentiles are established at the time the test is developed and published. It is frequently thought by educators that the national percentile is determined each year the test is administered, which is not the case.

PERCENTILE

A NUMBER THAT REPORTS THE RELATIVE RANK OF AN INDIVIDUAL WITHIN A GROUP, OR A GROUP WITHIN GROUPS.

- PERCENTILES ARE RANKS.
- NATIONAL PERCENTILES ARE ESTABLISHED AT THE TIME THE TEST IS NORMED. NOT A YEARLY PROCESS.

SLIDE 10

The two types of percentiles used most frequently are the national and local percentiles. The national percentile gives ranking of an individual with respect to a national reference group established at the time the test was published. The local percentile gives the ranking of an individual with respect to a defined local group and grade level tested. Local percentiles are generally developed each time a test is given.

PERCENTILESNATIONAL

GIVES RANKING OF AN INDIVIDUAL WITH RESPECT TO A NATIONAL REFERENCE GROUP (ON WHICH THE TEST WAS NORMED.)

LOCAL

GIVES RANKING OF AN INDIVIDUAL WITH RESPECT TO THE LOCAL GROUP AND GRADE LEVEL TESTED.

SLIDE 11

This transparency shows a distribution of raw scores from 1 to 28 for a given test. The individual percentile is the percentile most frequently found in test publisher's manuals and is used to show how an individual pupil ranks in relation to his/her peers in the national sample. Some test publishers also provide a group percentile which allows comparisons of class, school or district data with other classes, schools or districts involved in the norming sample. There is a tendency on the part of some educators to use the individual percentile distribution to determine the relative standing of school averages. This is not a proper use of the individual percentile table. Group percentiles are generally insensitive to major fluctuations in the extremes of the distribution and highly sensitive to minor variances in mean scores in the middle of the distribution.

<u>SCORE</u>	<u>INDIV %ILE</u>	<u>GROUP %ILE</u>	<u>SCORE</u>	<u>INDIV %ILE</u>	<u>GROUP %ILE</u>
1	1	1	23	99	99
2	1	1	24	99	99
3	2	1	25	99	99
4	4	1	26	99	99
5	7	1	27	99	99
6	11	1	28	99	99
7	18	4			
8	26	10			
9	36	23			
10	45	41			
11	52	63			
12	60	64			
13	67	85			
14	71	91			
15	77	96			
16	82	99			
17	85	99			
18	89	99			
19	92	99			
20	94	99			
21	96	99			
22	98	99			

SLIDE 12

This transparency shows a raw score scale, grade equivalent score scale, standard score scale and the percentile rank for a grade 10 group on a nationally standardized norm-referenced test. This particular test deals with language usage and structure. In this example, it doesn't make any difference whether the individual or the group has 31 or 54 items correct, the grade equivalent score is all the same - 13.6. There is no room for growth. When the individual or the group achieves a raw score of 31 the ceiling has been reached no matter how much the raw scores go up. Standard scores provide room for growth for either an individual or a group beyond the raw score of 31 where the standard score range will go from 648 to 999. Both the grade equivalent score and the percentile are influenced by a ceiling effect. A raw score of 42 or a raw score of 54 yields the same grade equivalent score and the same percentile. In addition to pointing out the unequal interval aspects of the grade equivalent and the percentile scale, this transparency dramatizes the problems that individual children or groups can have when they are at the extremes of the distribution. It takes much more change in raw score points to gain the same ground that you may cover with far fewer raw score points if you are at the average or near the average of the distribution.

LANGUAGE - USAGE AND STRUCTURE

<u>RAW</u> <u>SCORE</u>	<u>GE</u> <u>SCORE</u>	<u>STANDARD</u> <u>SCORE</u>	<u>GR.10</u> <u>%ILE</u>	<u>RAW</u> <u>SCORE</u>	<u>GE</u> <u>SCORE</u>	<u>STANDARD</u> <u>SCORE</u>	<u>GR.10</u> <u>%ILE</u>
1	.6	238	1	28	10.7	601	58
2	.6	240	1	29	11.9	617	64
3	.6	242	1	30	13.2	633	70
4	.6	246	1	31	13.6	648	75
5	.6	252	1	32	13.6	662	79
6	.6	258	1	33	13.6	676	83
7	.8	266	1	34	13.6	690	86
8	.9	275	1	35	13.6	703	89
9	1.1	285	1	36	13.6	716	91
10	1.3	296	1	37	13.6	729	93
11	1.5	309	1	38	13.6	742	95
12	1.8	322	1	39	13.6	755	96
13	2.1	336	1	40	13.6	768	97
14	2.4	352	1	41	13.6	782	98
15	2.7	368	1	42	13.6	796	99
16	3.1	385	2	43	13.6	811	99
17	3.4	402	3	44	13.6	827	99
18	4.0	420	5	45	13.6	843	99
19	4.6	438	7	46	13.6	861	99
20	5.3	457	10	47	13.6	879	99
21	6.0	475	14	48	13.6	899	99
22	6.8	494	18	49	13.6	918	99
23	7.4	513	24	50	13.6	938	99
24	8.1	531	30	51	13.6	958	99
25	8.9	549	37	52	13.6	977	99
26	9.7	567	44	53	13.6	994	99
27	10.2	584	51	54	13.6	999	99

SLIDE 13

Expectancy scores are sometimes referred to as anticipated achievement scores. Expectancy or anticipated scores are not always dealt with in all instruments because it requires the administration of both an achievement and an aptitude or intelligence test. This particular example shows the expectancy score as defined by CTB/McGraw-Hill's use of the Short Form Test of Academic Aptitude and its various achievement tests, such as the California Achievement Test or the Comprehensive Tests of Basic Skills. The expectancy or anticipated score is a statistical estimate (average) of all the pupils in the national reference group who are at the same age, grade, sex and academic aptitude as a pupil being tested. It is an additional way of comparing a pupil with a national reference group, only this group is more similar to the child being tested. Academic aptitude is used as one variable in predicting achievement based on the performance of pupils with similar characteristics. Primary variables used in predicting achievement are the four subtest scores on the Short Form Test of Academic Aptitude, age, grade and sex. The emphasis of this particular transparency is the predictive aspect of the aptitude measure, and how aptitude can be used to compare a child's progress with a more specific subgroup of the national sample.

EXPECTANCY SCORES

A STATISTICAL ESTIMATION OF THE EXTENT TO WHICH AN INDIVIDUAL PUPIL IS ACHIEVING IN ACCORDANCE WITH HIS/HER PEERS AT THE SAME AGE, GRADE, SEX AND ACADEMIC APTITUDE.

PRIMARY VARIABLES

AGE
GRADE
SEX
SFTAA SCORES ON
 VOCABULARY
 MEMORY
 ANALOGIES
 SEQUENCES

SLIDE 14

The remainder of this staff development program should be oriented toward actual test instruments or measurement techniques used in each school system. The particular examples used here are based on results from the Comprehensive Tests of Basic Skills.*

Standard error is one of the most important concepts associated with norm-referenced test interpretation. This is the basis for the development of the percentile bands which are used more frequently now by test publishers in their individual and group reports. The major ingredients that go into the derivation of standard error are reliability, standard deviation and number of items.

There are two types of standard error; measurement and of the mean. The standard error of measurement indicates how much an individual's score would vary if he or she were repeatedly tested with the same instrument and no learning occurred between test administrations. The standard error of the mean indicates how much the obtained average or mean from a group test is likely to vary from one group testing situation to another. The concept of standard error reflects the fact that tests are measures of behavior and behavior is not static. This variability is identified as measurement error. It is an extremely important concept in dealing with norm-referenced standardized tests because it doesn't allow specific categorization of pupils in terms of how they score on a test. There is always the possibility that a child could have a score other than what he or she actually received and that score could be well within the limit of the standard error of the instrument.

* CTB/McGraw-Hill, Monterey, California

STANDARD ERROROF MEASUREMENT

INDICATES HOW MUCH AN INDIVIDUAL'S SCORE WOULD VARY IF HE WERE REPEATEDLY TESTED WITH THE SAME INSTRUMENT AND NO LEARNING OCCURRED BETWEEN TEST ADMINISTRATIONS.

OF THE MEAN

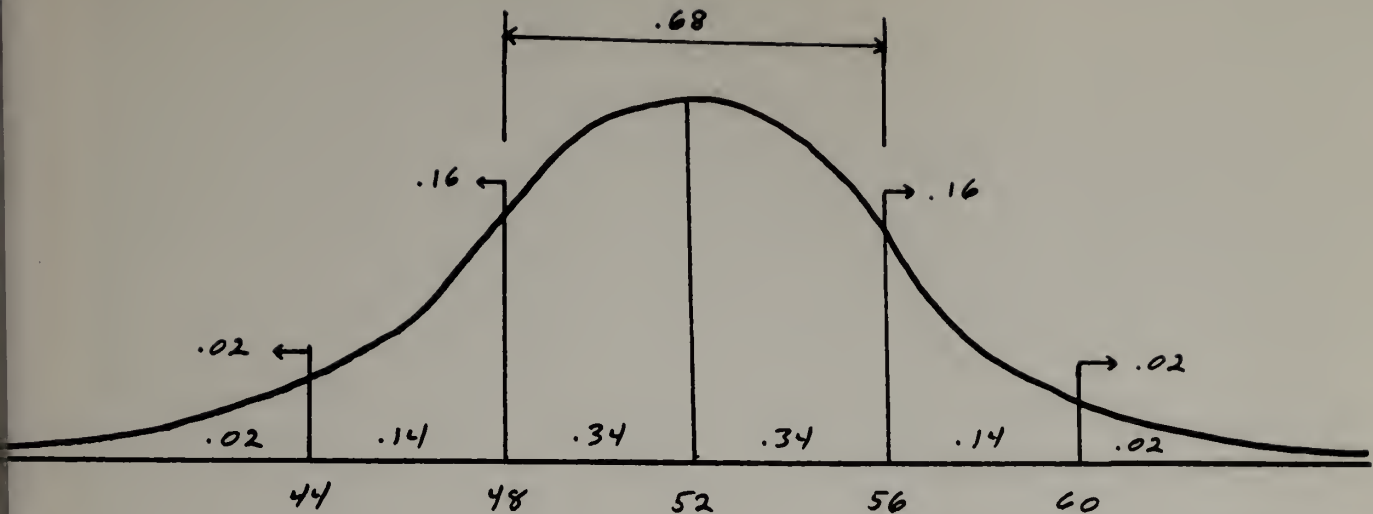
INDICATES HOW MUCH THE OBTAINED MEAN FROM A GROUP TEST IS LIKELY TO VARY FROM ONE GROUP TESTING SITUATION TO ANOTHER.

REFLECTS THE FACT THAT TESTS ARE A MEASURE OF BEHAVIOR AND BEHAVIOR IS NOT STATIC BUT VARIABLE. THIS VARIABILITY IS IDENTIFIED AS MEASUREMENT ERROR.

SLIDE 15

Standard error of measurement is similar in concept to standard deviation. Data from a test is described in terms of the central point (mean) and spread of scores around the mean (standard deviation). In the case of an individual pupil's score, the obtained score represents the middle of a possible distribution of scores if the child repeatedly took the test. The estimated degree of spread of all these possible scores around the child's "true" score is referred to as the Standard Error of Measurement. Standard Error of Measurement is based on the assumption that if an individual takes the same test several times without intervening instruction, the scores would differ to some degree. This variation will be less severe if pupils taking the test tend to score close to the average (low standard deviation) and if children tend to score the same way in repeated testing situations (high reliability).

The following example shows how the Standard Error of Measurement can be used to estimate the probability of a child's performance on a test. In this case, the teacher might wish to establish the probability level of Johnny having a passing score of 56 when his obtained score was 52.



- GIVEN: 1) SCIENCE MID-TERM EXAM
 2) STANDARD DEVIATION = 8.0
 3) RELIABILITY = .75
 4) JOHNNY'S SCORE = 52
 5) PASSING SCORE = 56

$$S.E.M. = 8 \sqrt{1.00 - .75} = 4.0$$

PROBABILITY OF JOHNNY'S SCORE BEING ABOVE 60 OR BELOW 44 IS
 .02 OR 1 CHANCE IN 50.

PROBABILITY OF JOHNNY'S SCORE BEING AT OR ABOVE PASSING IS
 .16 OR 1 CHANCE IN 6.

PROBABILITY OF JOHNNY'S SCORE BEING BETWEEN 48 AND 56 IS
 .68 OR 2 CHANCES OUT OF 3.

SLIDE 16

This transparency shows an example of standard error of measurement applied to raw scores, standard scores and percentiles. This example uses standard error in standard score units from the CTBS Form S Level 2 Grade 6 Vocabulary subtest. The standard error of measurement is 23 score units. A pupil with 35 items correct on this particular test would have a converted standard score of 541. A standard score of 541 is at the 80th percentile. However, if the standard error of 23 points is both added to and subtracted from the score of 541, the results is a range of from 518 to 564. This is interpreted by saying that if this child was repeatedly tested with this same test without intervening instruction, 68% of the time the child's score would not be less than 518 or more than 564. This range of scores is referred to as a "percentile band." The percentile band contains the range of percentiles where a child with an obtained score could possibly be performing if he were retested again without intervening instruction. The degree of confidence one could place on the location of the "true" score this child would obtain can be increased to 95% by doubling the standard error to 46 score units. However, most test publishers hold to the 68% confidence interval, since it encompasses the range of 1 standard error of measurement in either direction of the obtained score.

GRADE 6 VOCABULARYCTBS-S LEVEL 2

STANDARD ERROR OF MEASUREMENT = 23 STANDARD SCORE UNITS

68% OF THE TIME THE SCORE A STUDENT OBTAINS ON A TEST WILL NOT VARY FROM THE "TRUE" SCORE BY MORE THAN 1 SEM IN EITHER DIRECTION.

95% OF THE TIME THE SCORE A STUDENT OBTAINS ON A TEST WILL NOT VARY FROM THE "TRUE" SCORE BY MORE THAN 2 SEM IN EITHER DIRECTION.

<u>OBTAINED STANDARD SCORE</u>		<u>RAW SCORE</u>	<u>NATIONAL PERCENTILE</u>	
495	-2 SEM	31	64	
518	-1 SEM	33	71	
→ 541		35	80	
564	+1 SEM	36	35	
587	+2 SEM	38	94	

* 68% CONFIDENCE INTERVAL

SLIDE 17

The previous example of the standard score of 541 in Vocabulary is shown on the following Individual Test Record. The resulting national percentile achieved here is 80. The percentile band to the right of this profile shows the range in x's of the child's performance using the standard error of measurement. The application of standard error toward determining significant differences between test performance (overlapping vs. non-overlapping bands), and the effect test reliability and number of items has on the width of the bands is demonstrated. Also the need to focus on item performance characteristics by skill level to determine possible deficiencies not uncovered by the display of scores alone is shown. For example, the relatively good scores in Language Expression and noticeable lack of correct responses in the process/content skill classification of Interpretation/Syntactical Relationships.

The scores and data on the Individual Test Record are defined as follows:

RS - Raw Score

The raw score is the number of items a child answered correctly on the test. It has no meaning by itself, but is used in developing other scores described below.

OSS - Obtained Scale Score

The Scale Score is a three-digit score on a scale between 000 and 999. The score a child receives in this column represents the obtained Scale Score converted from the number of items correct (raw score) for each test. The average Scale Score will vary depending on the grade and level of the test and should not be interpreted as being 500 for each grade and test

level. The Scale Score provides a way of measuring growth between successive testings that cannot be obtained from other scores.

AASS - Anticipated Achievement Scale Score

The column headed AASS shows scores on each test obtained by students with similar age, grade in school, sex and academic aptitude. This score is provided only if the child took both CTBS and a test of academic aptitude. It can be used as a meaningful aid in further identifying skill strengths or weaknesses only if it is significantly higher or lower than the obtained scale score. When this is the case, the difference between the SS and AASS is printed in the Difference (DIFF) column.

DIFF - Difference

The column headed DIFF records the differences between the Obtained Scale Score (OSS) and the Anticipated Achievement Scale Score (AASS) only when these differences are large enough to be important. If the obtained scale score that a child received is significantly higher than the AASS, the difference printed has a plus sign. This can be interpreted as meaning that the child performed significantly better on this test than other students with the same age, grade, sex, and academic aptitude. A minus sign indicates that the child has scored significantly lower than other students with the same age, grade, sex, and academic aptitude.

NP - National Percentile Rank

LP - Local Percentile Rank

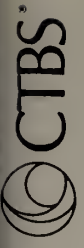
Percentile rank shows the percentage of students in the national sample or local group that received a lower score than a particular child. These

scores tell how this child ranks either nationally or locally on these tests with respect to other children at the same grade level. For example, a 7th grade student with a National Percentile Rank of 48 in Spelling means that 48% of the 7th grade students in the national sample received a lower Spelling score. A Percentile Rank score of 50 is average for a child's particular grade on both the national and local scale. There may be differences between these two scales in a school district since the school population may or may not be similar to the national sample.

Percentile Rank Chart

The Percentile Rank Chart in the upper right-hand portion of this report gives a graphic picture of a child's test scores. This chart uses only National Percentile Rank information. Because test scores are not exact measures of a student's achievement, the row of x's for each test shows the range of potential achievement within which a child's score is most likely to fall.

The item analysis is an important feature of this particular report. Each subtest is broken down into the process and content skill classifications showing the item numbers and the pupil's responses. A plus (+) indicates a correct response, a minus (-) indicates an incorrect response, and a blank indicates the child omitted the item. This analysis gives the teacher considerable information and can be used to diagnose potential skill deficiencies.



NAME **A** TEACHER **A4E2** BATCH **A4E2**
 SCHOOL **026** GROUP **026** RUN DATE **10/09/76**
 CITY **06.C** GRADE **06.C**
 DATE OF TESTING **05/76** NATIONAL PERCENTILE

TEST	RS	CS	AASS	DIFF	NP	LP	1	2	3	4	5	6	7	8	9	10	11	12	
READING VOCABULARY	35	541	502		80	59													
READING COMPREHENSION	32	504	530		56	34													
TOTAL READING	67	510	506		68	44													
SPELLING	35	477	496		48	44													
LANGUAGE MECHANICS	15	507	519		56	42													
LANGUAGE EXPRESSION	27	550	537		72	46													
TOTAL LANGUAGE	81	498	503		60	42													
MATHEMATICS COMPUTATION	34	442	469		47	33													
MATHEMATICS CONCEPTS	16	454	458		50	26													
MATHEMATICS APPLICATIONS	5	385	506	-121	20	5													
TOTAL MATHEMATICS	55	427	476		38	19													
TOTAL BATTERY	207	455	481		55	32													
REFERENCE SKILLS	12	464	526		40	15													
SCIENCE																			
SOCIAL STUDIES	22	495	519		56	23													

READING VOCABULARY		LANGUAGE MECHANICS		LANGUAGE EXPRESSION		INTERPRETATION	
RECOGNITION/APPLICATION	ANALYSIS	RECOGNITION/APPLICATION	CAPITALIZATION	RECOGNITION/APPLICATION	USAGE	TRANSLATION	DICTION
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40	+	+	+	+	+	+	+
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20	+	+	+	+	+	+	+
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20	+	+	+	+	+	+	+

READING COMPREHENSION		ANALYSIS	
RECOGNITION/APPLICATION	INTERPRETATION	STRUCTURE/STYLE	CONCLUSIONS
LITERAL RECALL	MAIN IDEA	CONTEXT CLUES	WORD RECOGNITION
6 13 16 21 22 26 33 34 8 20 23 36 41 45 7 9 14 15 28 2 12 24 38 44 19 30 32 39 1 10 11 17 31 37 42 43 3 4 5 18 25 27 29 35 40	+	+	+

MATHEMATICS COMPUTATION		MATHEMATICS CONCEPTS	
APPLICATION	MULTIPLICATION	INTERPRETATION/ANALYSIS	PROBLEM SOLVING
ADDITION	SUBTRACTION	MEASURE MEASUREMENT	NUMBER SYSTEMS
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48	+	+	+

REFERENCE SKILLS		SCIENCE	
RECOGNITION/APPLICATION	LIBRARY USE	RECOGNITION	CLASSIFICATION
DICTIONARY SKILLS	LIBRARY USE	HECOGNITION	CLASSIFICATION
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 1 2 3 4 5 6 7 8 9 10 11 12 13	+	+	+

SOCIAL STUDIES		MATHEMATICS APPLICATIONS	
RECOGNITION	TRANSLATION	INTERPRETATION	PROBLEM SOLVING
RECOGNITION	TRANSLATION	MEASURE MEASUREMENT	NUMBER SYSTEMS
1 3 18 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55	+	+	+

From the Comprehensive Tests of Basic Skills, Copyright © 1973 by McGraw-Hill, Inc. All Rights Reserved. Printed in the U.S.A.

SLIDE 18

Standard error can also be applied to the interpretation of I.Q. data. This transparency dramatizes problems associated with I.Q. interpretation, in an attempt to reduce rigid classifications of pupils based on the data. Also shown are different I.Q. scores a child could have if he performed at the same level on each of three I.Q. tests. I.Q. scores alone are meaningless unless the score scale of the test, particularly the mean and standard deviation, are known and understood by the user. Educators are urged to examine trends in individual I.Q. scores over time and concentrate on the national percentile as the primary index of scholastic aptitude.

PROBLEMS WITH I.Q. INTERPRETATION

GIVEN: PUPIL WITH A TESTED I.Q. OF 100
 STANDARD ERROR: 5

<u>PROBABILITY</u>	<u>I.Q. SCORE RANGE</u>
68 CHANCES OUT OF 100	95 - 105
95 CHANCES OUT OF 100	90 - 110
99 CHANCES OUT OF 100	85 - 115

IN STANFORD-BINET TERMINOLOGY A RANGE OF 85-115 TRANSLATES TO A RANGE OF "DULL NORMAL" TO "ABOVE AVERAGE."

DIFFERENT TESTS WILL YIELD DIFFERENT SCORES FOR THE SAME LEVEL OF PERFORMANCE:

<u>TEST</u>	<u>MEAN</u>	<u>STANDARD DEVIATION</u>	<u>SCORE PLUS 3 STANDARD DEVIATIONS ABOVE THE MEAN</u>
KUHLMANN-ANDERSON, 6TH ED.	100	12	136
STANFORD-BINET	100	16	148
ARMY GENERAL CLASSIFI- CATION TEST	100	20	160

SLIDE 19-20

In addition to reports explaining individual pupil performance, most norm-referenced test publishers provide group reports. These group reports are generally class lists focusing on scores of pupils in a designated class or instructional grouping. This example of a Class Record Sheet from the CTBS is a partial profile of the subtests and only one pupil is listed. The areas stressed here are patterns shown in the difference between Language and Non-Language aptitude scores, and Reading Comprehension is amplified with the following transparency.

CLASS RECORD SHEET

NAME	APTITUDE			READING				LANGUAGE			
	LANG	NON-LANG	TOTAL	VOCABULARY	COMPREH	TOTAL	MECHANICS	EXPRESSION			
	1	2	3	1	2	3	1	2			
SMITH SAM	127 91	101 41	113 68	456 81 462	543 92 473 +70	486 88 458	618 98 481 ++	315 3 392	3 -77		

92

THE NATIONAL PERCENTILE - 92% OF THE NATIONAL SAMPLE HAD SCORES BELOW THIS OBTAINED STANDARD SCORE OF 543.

+70

THE DIFFERENCE BETWEEN 543 - 473 = 70 WHICH FOR THIS SUBTEST AND THIS CHILD'S APTITUDE CHARACTERISTICS, AGE, GRADE AND SEX IS CONSIDERED STATISTICALLY SIGNIFICANT.

543

THE STANDARD SCORE THE CHILD RECEIVED ON THE TEST.

473

THE AVERAGE STANDARD SCORE RECEIVED BY CHILDREN WHO HAD THE SAME APTITUDE SCORES AS THIS CHILD AND WHO WERE THE SAME AGE, GRADE AND SEX AS THIS CHILD.

SLIDE 21-22

The most meaningful and useful instructional information from test results is obtained from item analysis. This information is particularly valuable if the item response data represents pupils' performance on skills measured by the test. Most test publishers offer types of item analysis information, which may range from very simple tables of percentages of correct responses per item to rather sophisticated item/skill diagnostic profiles. The following two examples show item analysis profiles both for a particular instructional group (Group Right Response Record) and by grade level (Right Response Summary).

GROUP RIGHT RESPONSE RECORD
 TEACHER JONES LC
 SCHOOL CENTRAL
 CITY ANYTOWN STATE CA
 GRADE ID 05.1-001-001
 CTB ID 9395-001-001
 DATE OF TESTING 10/73
 RUN DATE 12/19/73
 NUMBER OF CASES 12
 TEST CTRS FORM S LEVEL 2
 COPYRIGHT (C) 1973
 BY MCGRAW-HILL, INC.

TEST	CONCEPTS	STUDENT COUNT	PCT RIGHT	LOC NAT
MATHEMATICS	CONCEPTS	12		
TEST SECTION	STUDENT COUNT =	12		
RECOGNITION	RELATIONSHIPS	58	50	
10	EQUILATERAL TRIANGLE	83	85	
13	CLOSED CURVE	58	51	
14	DIAMETER	42	48	
15	PARALLELOGRAM RECOGNITION			
TRANSLATION				
GRAPHS		100	78	
03	CIRCLE GRAPH	75	74	
04	LINE GRAPH	33	36	
05	ORDERED PAIR OF NOS			
INTERPRETATION		58	33	
NUMBER SYSTEMS/PROPERTIES		100	79	
06	GREATEST VALUE - FRACT	58	51	
07	NUMERAL IOENT - FOUR-DIG	67	25	
08	PLACE VALUE - DECIMAL NO	75	74	
09	LEAST VALUE - FRACT			
16	FRACTIONAL PART	92	81	
SETS		75	49	
01	CARDINALITY			
02	FRACTIONAL PART OF A SET			
MEASUREMENT		67	30	
12	UNITS OF AREA	42	71	
19	CONVERT AND COMPARE IN AND FT	33	41	
20	CONVERT YDS TO FT - OPER NEEDED			
MATHEMATICAL SENTENCES		42	53	
22	INEQUALITIES - SYMBOLS			
PROBLEM SOLVING		8	42	
17	WORD PROB - (AVERAGE)-OPER NEEDED	25	59	
18	WORD PROB - (MULTI)-OPER NEEDED	8	59	
23	WORD PROB - (MULTI)-OPER NEEDED	8	58	
25	WORD PROB - (MULTI)-OPER NEEDED			
ANALYSIS		42	49	
MEASUREMENT				
11	CONVERT GAL TO QT	17	70	
PROBLEM SOLVING		17	33	
21	WORD PROB (AVERAGE)-INFO NEEDED			
24	WORD PROB - SET UP, SOLVE			

RIGHT RESPONSE SUMMARY

	NATIONAL REFERENCE GROUP						DIFF
	4.7		5.1*		5.7		
	OMITS PCT	WRONG PCT	RIGHT PCT	RIGHT PCT	RIGHT PCT	RIGHT PCT	
MATHEMATICS CONCEPTS	THE NUMBER OF CASES IN THIS SECTION IS 85.						
RECOGNITION	6	33	60	58	60	63	0
GEOMETRIC RELATIONSHIPS	6	45	49	58	60	63	-11
10 EQUILATERAL TRIANGLE	28	40	32	48	50	54	-18
13 CLOSED CURVE	0	17	83	87	89	91	- 6
14 DIAMETER	17	25	58	26	33	43	25
.							
.							
.							
.							
↓							

* THIS VALUE REPRESENTS A LINEAR INTERPOLATION BETWEEN THE TWO OUTER VALUES.

SLIDE 23

A frequency distribution, either computer or teacher made, can provide useful information as to the position pupils or groups occupy in a score distribution. In this example the number of pupils scoring in each range of scores is shown in the frequency column. The average (mean) score position for both Vocabulary and Comprehension is shown by the horizontal line. In the Vocabulary subtest there are two distinct scoring groups. There are 10 children scoring relatively low in the distribution and 15 scoring above the class average. In Comprehension, the group's scores are widely distributed throughout the scale with no apparent clustering at any particular score range. This information could have significance for classroom instructional grouping practices.

FREQUENCY DISTRIBUTION

	<u>READING VOCABULARY FREQUENCY</u>	<u>READING COMPREHENSION FREQUENCY</u>
601 - 607		1
594 - 600		
587 - 593		
580 - 586		1
573 - 579		
566 - 572		1
559 - 565		
552 - 558		1
545 - 551		
538 - 544		1
531 - 537		
524 - 530	2	1
517 - 523	2	1
510 - 516	3	1
503 - 509	2	1
496 - 502	2	1
489 - 495	4	
482 - 488		1
475 - 481		1
468 - 474	<hr/>	<hr/>
461 - 467		2
454 - 460		
447 - 453		2
440 - 446		
433 - 439	4	1
426 - 432	3	
419 - 425	3	2
412 - 418		1
405 - 411		2
398 - 404		1
391 - 397		
384 - 390		1
377 - 383		1

SLIDE 24

If proper scores are used, growth patterns can be assessed with norm-referenced tests. This transparency shows how a child's math data has been plotted on a growth chart for four years. The growth pattern is consistent with the national sample average from grade 4 through 6; however, the pattern from grade 6 to 7 shows a lack of skill growth.

Growth Chart

EXPANDED STANDARD SCORES (SCALE SCORES)



TOTAL MATHEMATICS

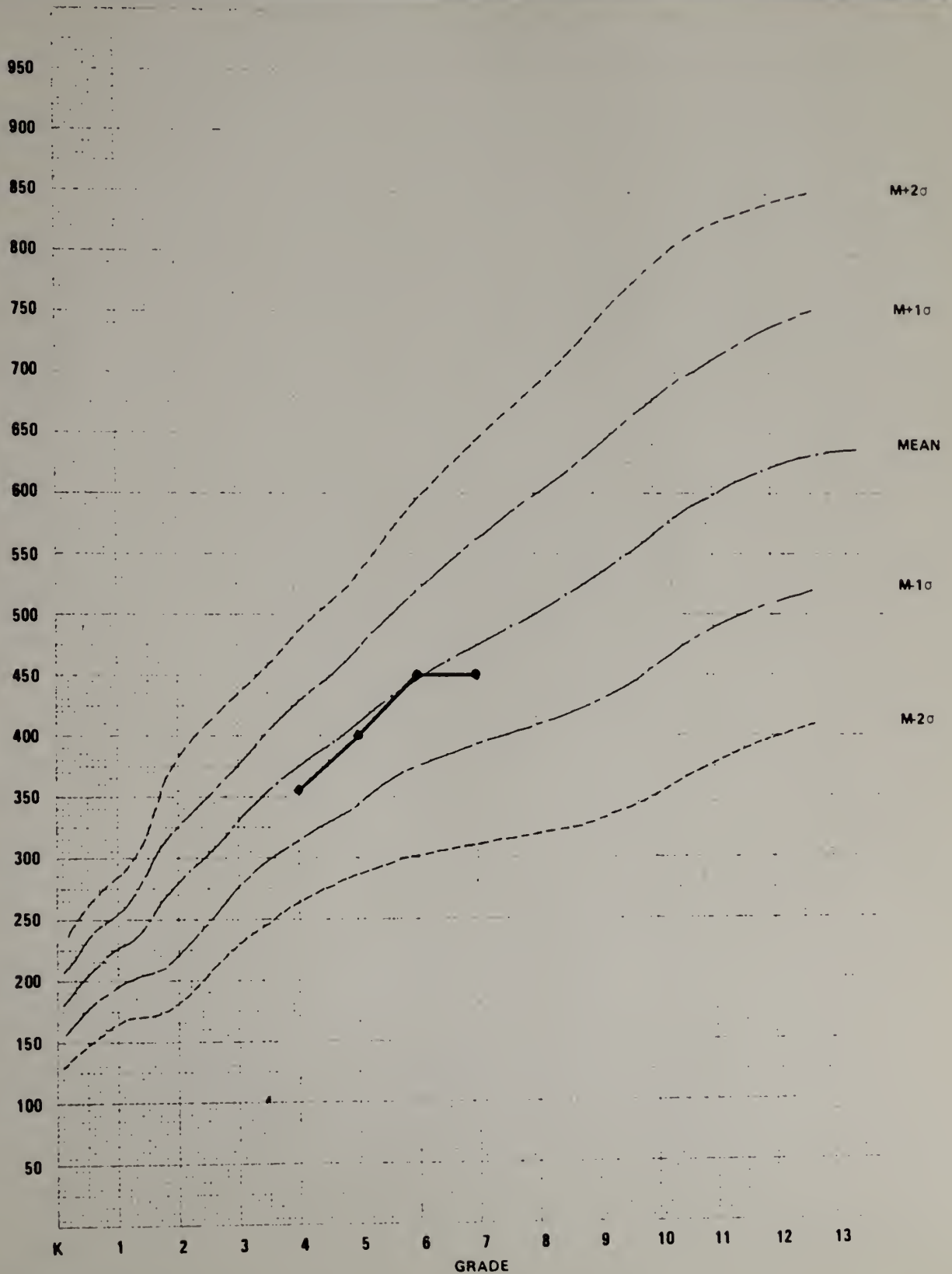
LEVELS A, B, C, 1, 2, 3, 4

UNDERGARTEN STREAM

Comprehensive Tests of Basic Skills

Expanded Edition

EXPANDED STANDARD SCORE (SCALE SCORE)



SLIDE 25-26

Effective use of test data requires knowledge by the user of what the test is designed to measure. Many teachers are asked to administer a test and use its results without fully understanding the rationale or intent of the instrument. This is particularly true in the case of town-wide or state-wide mandated testing programs. The following two transparencies show general objectives of the CTBS and the item numbers that measure the specific process/content skill classifications. Teachers and administrators are urged to become aware and examine these documents to have a better understanding of the overall purpose of the testing instrument. This understanding may tend to encourage some to rely more on skill related responses rather than the usual simplistic focus on scores.

Slides 25 and 26 are modified forms taken from the Comprehensive Tests of Basic Skills Test Coordinators Handbook (Preliminary Edition). Reproduced by permission of the publisher, CTB/McGraw-Hill, Monterey, CA 93940. Copyright © 1974 by McGraw-Hill, Inc. All Rights Reserved. Printed in the U.S.A.

CONTENT CATEGORIES FOR LANGUAGE, LEVELS 1-4

TEST 3 - SPELLING

TEST 4 - LANGUAGE MECHANICS

TEST 5 - LANGUAGE EXPRESSION

RECALL OF RULE (3)

RECOGNIZE CORRECTLY AND INCORRECTLY SPELLED WORDS.

PUNCTUATION (4)

SELECT THE PUNCTUATION MARK NEEDED IN A GIVEN SENTENCE.

CAPITALIZATION (4)

SELECT THE SEGMENT OF A GIVEN SENTENCE THAT CONTAINS AN ERROR IN CAPITALIZATION.

USAGE (5)

SELECT THE GRAMMATICAL FORM REQUIRED TO COMPLETE A GIVEN SENTENCE.

CONTEXT CLUES (5)

USE CONTEXT CLUES TO DECIDE WHETHER OR NOT ONE OF A PAIR OF HOMONYMS OR EASILY CONFUSED WORDS IS USED CORRECTLY IN A GIVEN SENTENCE.

DICTION (5)

USE CONTEXT CLUES TO SELECT THE WORD THAT BEST COMPLETES THE SENTENCE IN TERMS OF THE IDEA BEING EXPRESSED.

SYNTACTICAL RELATIONSHIPS (5)

UNDERSTAND THE INTERRELATIONSHIPS OF SENTENCE STRUCTURE AND SEMANTICS.

ORGANIZATION (5)

SELECT THE CONNECTIVE (CONJUNCTION OR TRANSITION WORD) THAT SHOWS THE RELATIONSHIP IN THOUGHT BETWEEN TWO SENTENCES IN A GIVEN PARAGRAPH, SELECT THE SENTENCE THAT SHOULD COME FIRST IN A PARAGRAPH OF FOUR SENTENCES IN SCRAMBLED ORDER, OR PUT FOUR SENTENCES THAT ARE GIVEN IN SCRAMBLED ORDER IN THE SEQUENCE THAT WOULD BEST EXPRESS THE FLOW OF THOUGHT IN THE PARAGRAPH.

ITEM CLASSIFICATION FOR LANGUAGE, LEVEL 2

TEST 3 - SPELLING TEST 4 - LANGUAGE MECHANICS TEST 5 - LANGUAGE EXPRESSION

Content Process	TEST 3		TEST 4		TEST 5			
	Recall of Rule	Context Clues	Punctua- tion	Capitali- zation	Usage	Diction	Syntacti- cal Rela- tionships	Organization Sequencing
Recognition/ Application	1-5,7-30 32-37,39 40,41,44 45,46,50		1-10	11-20	21-30			
Translation		6,31,38,42 43,47,48,49				31-44		
Interpretation							45-49	
Analysis								50-55

SLIDE 27

A serious problem confronting most educators, particularly educational administrators, is how to display district data from norm-referenced standardized tests. The power of a norm-referenced standardized achievement test is its ability to compare a pupil with a reference group, and provide a gross measure of possible skill deficiencies for both individuals and groups. However, the push for accountability has placed these test results into a role they were not designed to fulfill. A general lack of understanding on the part of users both inside and outside of the academic environment further complicates the issue. Group data or district-wide data have value if properly presented. The common practice for some school systems is to present average (mean) scores, usually in the form of grade equivalents, by grade level and subtest categories (i.e. Reading, Language, Math, etc.). Some problems associated with using grade equivalent scores this way have been discussed. Problems of using average scores as a sole index of achievement have also been discussed. However, the public has seen and expects to see average scores by skill area. There is an attempt on the part of some school systems to use standard scores instead of grade equivalent scores in presenting summary data by district. This is a more legitimate procedure though the average of any type of score is very limited in what it can provide toward effective decision making. Low performance in some tested skill areas at a specific time of the year may not necessarily be indicative of failing educational practices. Different instructional approaches with children at certain times, which may not coincide with the content of a subtest, may result in immediate low performance but pay off with greater future gains. Whatever method of displaying data is used, it should not involve the display of school by school comparisons within a district or comparisons between districts.

This is particularly true in a community where schools may differ with respect to socio-economic characteristics. Comparisons between districts are meaningful only if the public understands the demographic characteristics and the educational differences that may exist.

PUBLIC DISPLAYS OF GROUP DATA

GRAPHIC DISPLAYS OF AVERAGES

LONGITUDINAL GRAPHIC DISPLAYS OF GROWTH PATTERNS

GRAPHIC DISPLAYS OF SPREAD AND SHAPES OF SCORE
DISTRIBUTIONS

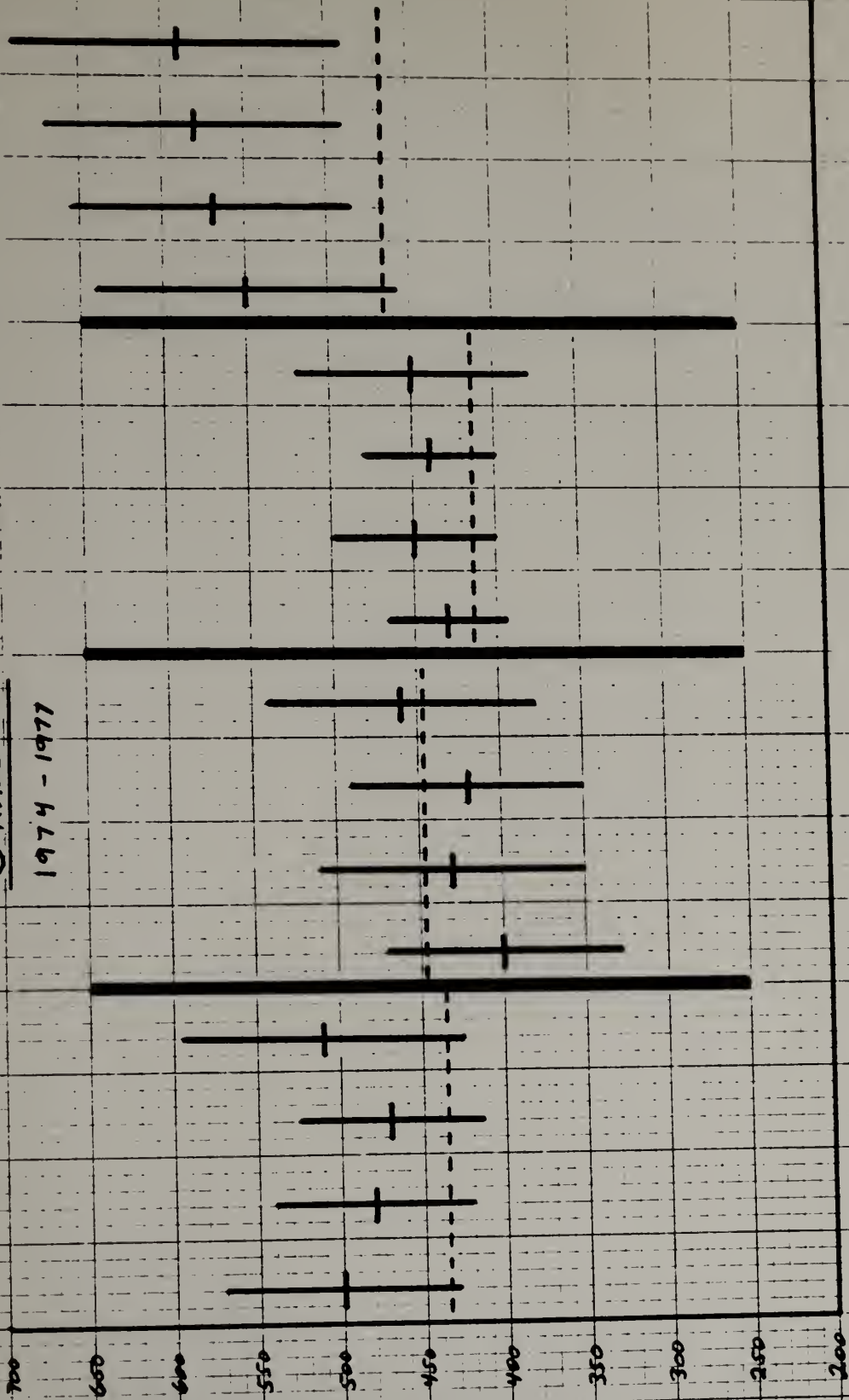
IDENTIFICATION OF SKILL DEFICIENCIES AND
INSTRUCTIONAL PLANS TO CORRECT THEM

SLIDE 28

One way of displaying district data is by showing the performance of pupils by grade level and comparing these data with the national average. This example shows the performance of grade 5 pupils in the district for four consecutive years. These are not the same pupils moving in time but are different classes. The dotted line reflects the national average (mean) for each one of the subtests. The short horizontal lines show the district averages (means) over four years. The vertical lines show one standard deviation or either side of the mean - the range where approximately 68% of the pupils are performing. This approach shows the average which the public seems to demand, but also shows the range of the majority of scores.

GRADE 5

1974 - 1977



REFERENCES

MATH

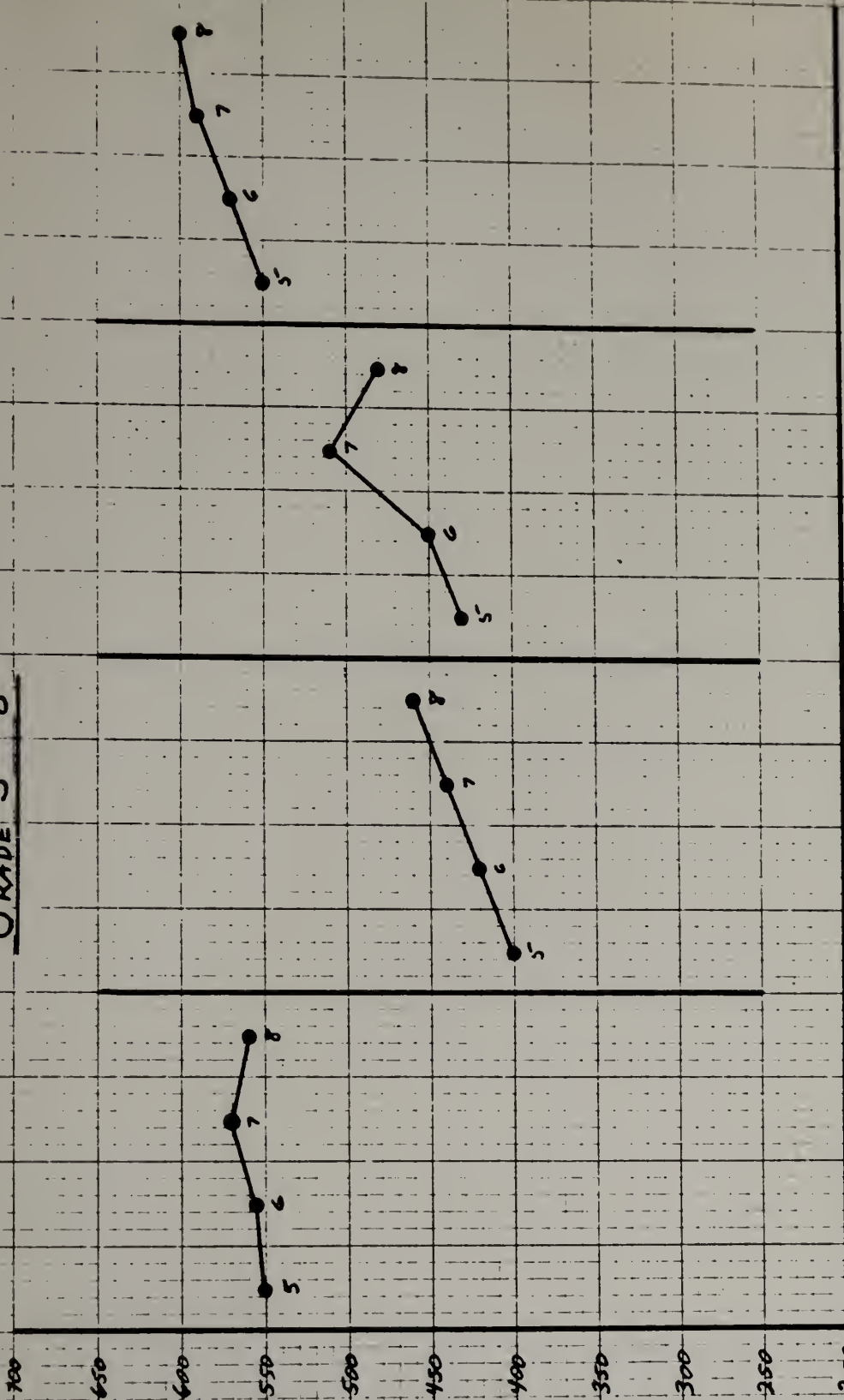
LANGUAGE

READING

SLIDE 29

Another procedure is shown in this example where the same pupils are followed through four consecutive years of testing. The purpose is to gain information regarding growth patterns and draw attention to areas of lack of growth. Areas of possible concern here are in reading and math between grades 7 and 8.

GRADE 5 - 8



REFERENCES

MATH

LANGUAGE

READING

SLIDE 30

An ideal method of presenting group or district data is when all three characteristics of the distribution can be displayed. The following is an example from the new California Achievement Test* where the distribution of a district's data is superimposed over the normal curve (normative distribution). The mean and standard deviation are shown below the curve for both the standard scores (SS) and the raw scores (RS). In this example, the district is shown the central point of their data (mean), the spread of the scores (standard deviation), and the shape of the distribution as compared with the national reference group.

* California Achievement Test, 1977: CTB/McGraw-Hill, Monterey, California.

Percentage per
Scale Score Unit

1.0%

0.8

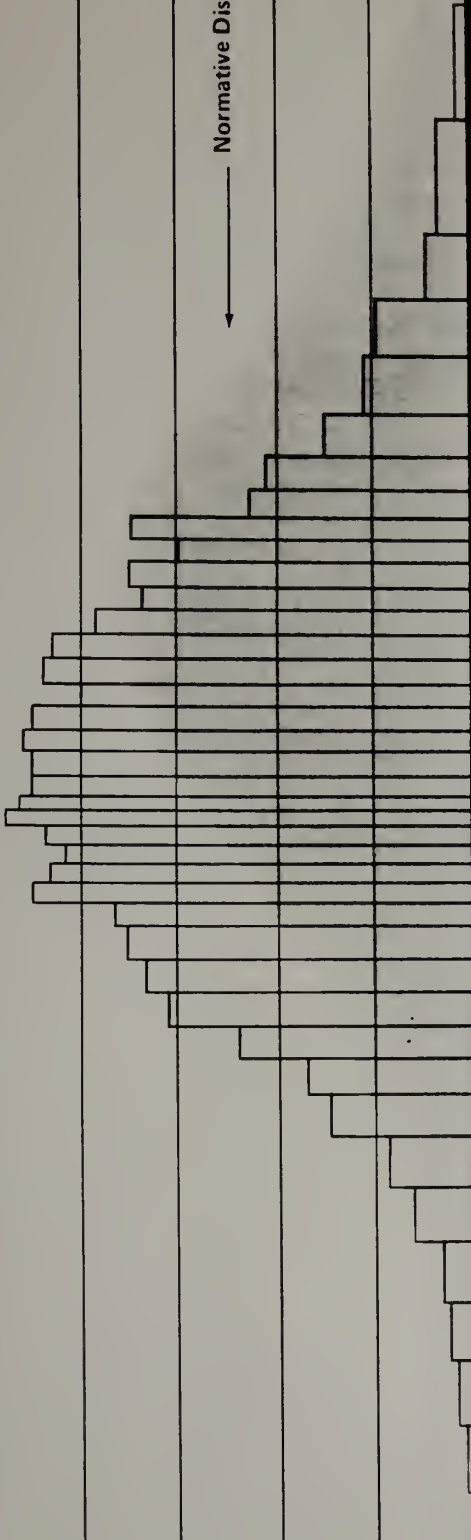
0.6

0.4

0.2

0

Normative Distribution



SS

RS

Difference between obtained and
normative distribution

+0.2

-0.2



SLIDE 31

One trend in analyzing and presenting summary data from tests is to provide schools or districts with narrative reports. An example of such an effort is the School Needs Assessment Profile (SNAP) developed by CTB/McGraw-Hill. The following is a copy of a SNAP report for a school where skill areas that meet specified criteria as being deficient are stated and prioritized for correction by the staff. This is a particularly valuable service to school systems that do not have the in-house capability to analyze in great detail the various reports that are available.

ELMVIEW SCHOOL

The following outline presents behavioral objectives that appear to need attention.

READING (Priority 3)

Grade 5

I. Vocabulary

Recall of Synonym

Given a word in a short phrase and a choice of four words, the student will choose synonyms for adjectives.

II. Comprehension

Words in Context

The student will choose the best meaning for a word presented in the context of a reading passage.

Author Technique

The student will identify the methods used by an author to present a subject, including stating facts, asking questions, giving opinions, and telling stories.

LANGUAGE (Priority 1)

Grade 5

I. Spelling

Words in Context

Given a sentence with a word underlined, the student will indicate whether the underlined word is spelled correctly. Misspelled words involved double consonants, letter reversal, and silent letter.

II. Mechanics

Punctuation

The student will select the punctuation mark required in a given sentence.

Capitalization

Given a sentence divided into sections, the student will identify the section where a capital letter is required.

