1987

Achievement directed leadership

Douglas A. Gentile
Iowa State University

INFORMATION TO USERS

While the most advanced technology has been used to photograph and reproduce this manuscript, the quality of the reproduction is heavily dependent upon the quality of the material submitted. For example:

- Manuscript pages may have indistinct print. In such cases, the best available copy has been filmed.

- Manuscripts may not always be complete. In such cases, a note will indicate that it is not possible to obtain missing pages.

- Copyrighted material may have been removed from the manuscript. In such cases, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, and charts) are photographed by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps. Each oversize page is also filmed as one exposure and is available, for an additional charge, as a standard 35mm slide or as a 17"x 23" black and white photographic print.

Most photographs reproduce acceptably on positive microfilm or microfiche but lack the clarity on xerographic copies made from the microfilm. For an additional charge, 35mm slides of 6"x 9" black and white photographic prints are available for any photographs or illustrations that cannot be reproduced satisfactorily by xerography.

Gentile, Douglas A.

ACHIEVEMENT DIRECTED LEADERSHIP

Iowa State University PH.D. 1987

PLEASE NOTE:

In all cases this material has been filmed in the best possible way from the available copy. Problems encountered with this document have been identified here with a check mark __✓__.

1. Glossy photographs or pages _____
2. Colored illustrations, paper or print _____
3. Photographs with dark background _____
4. Illustrations are poor copy __✓__
5. Pages with black marks, not original copy __✓__
6. Print shows through as there is text on both sides of page _____
7. Indistinct, broken or small print on several pages __✓__
8. Print exceeds margin requirements __✓__
9. Tightly bound copy with print lost in spine _____
10. Computer printout pages with indistinct print _____
11. Page(s) _____ lacking when material received, and not available from school or author.
12. Page(s) _____ seem to be missing in numbering only as text follows.
13. Two pages numbered _____. Text follows.
14. Curling and wrinkled pages _____
15. Dissertation contains pages with print at a slant, filmed as received __✓__
16. Other _____

_____

_____

University
Microfilms
International

Achievement directed leadership

by

Douglas A. Gentile

A Dissertation Submitted to the

Graduate Faculty in Partial Fulfillment of the

Requirements for the Degree of

DOCTOR OF PHILOSOPHY

Department: Professional Studies in Education
Major: Education (Educational Administration)

Approved:

Signature was redacted for privacy.

In Charge of Major Work

Signature was redacted for privacy.

For the Major Department

Signature was redacted for privacy.

For the Graduate College

Iowa State University
Ames, Iowa

1987

TABLE OF CONTENTS

LIST OF TABLES

CHAPTER I. ACHIEVEMENT DIRECTED LEADERSHIP

Introduction

One contemporary trend in educational research is to study the profiles of effective school organizations in the hope that critical elements which contribute to their success can be identified. The logical assumption is that once identified, these practices can be instituted by other schools in an effort to elevate student achievement. Research of this genre evolved largely from the findings of the Coleman Report (1966), and the Jencks study (1972), which suggested that individual schools do not make a difference; a student's achievement is primarily a function of family background. Subsequently, several studies have been undertaken to test the accuracy of this finding (Sweeney, 1981; Purkey and Smith, 1982), the results of which seem to indicate that schools do make a difference and that important characteristics of these schools can be pinpointed.

Effective schools studies have been pursued by research teams led by Lezotte et al., 1974; Edmonds and Fredrichsen, 1979; Brookover, 1979; Austin, 1978; and Rutter and Mortimer, 1979; and, although each investigation has accentuated varying aspects of the schools under study, they seem to be in accordance on certain points. As a result, a review of effective schools research yields five correlates which are identified with effective school organizations. They are: (1) positive school climate, (2) strong leadership, (3) high expectations for student achievement and behavior, (4) clearly defined goals and a sense of purpose, and finally, (5) systematic monitoring of student achievement. Whereas all these correlates seem to be identified with school

effectiveness, the final element (systematic monitoring of student achievement) represents a double bonus, since it also provides a quantitative indicator of student learning. This indicator may be used to judge the efficacy of the instructional program as well as the appropriateness of the content material. As Hudgins and Phye (1983) state: "Scientific principles of measurement and evaluation are applied in the classroom in order to provide an objective basis for judging the progress of a student's learning, the effectiveness of teacher's methods, and the functional effects of school programs and policies."

In April of 1983, the report generated from an 18-month study produced by the National Committee on Excellence in Education suggests that a system of "standardized tests of achievement should be administered at major transition points from one level of schooling to another...." Previous to that conclusion, Ahmann and Glock (1981) stated that "...evaluation of the degree to which educational goals have been achieved is a basic part of teaching and concerns everyone associated with the school." Finally, by issuing a direct charge to the classroom teacher, Popham (1981) warns, "An educator who fails to become conversant with the current considerations in educational measurement is an educator destined to deal unsatisfactorily with a host of educational problems."

Clearly, these authorities espouse the importance of student achievement measures in the classroom. The same sentiments are described in the works of Robert Ebel (1965), Gordon Cawelti (1984), and Anne Anastasi (1976). The problem facing classroom teachers and principals is what type of tests to use and how to utilize the results in a meaningful

way. The choices in paper and pencil tests range from norm-referenced, standardized, published tests to criterion-referenced, teacher-made measures. For the purpose of this study, the researcher has chosen to focus on teacher-made, criterion-referenced tests, since they represent the most accurate measure of student learning with reference to a specific domain of knowledge such as a normal classroom curriculum (Zavarella, 1980; McKenna, 1977; Glaser, 1963; Popham, 1978). While not discounting the usefulness of norm-referenced tests for determining comparative achievement among students, the fact remains that when teaching for specific subject mastery, the criterion-referenced test is the better indicator of success (Carroll, 1963; Bloom, 1976).

Statement of the Problem

The implications of a systematic methodology for monitoring student progress are that the individual teachers apply measurement theory correctly in their classrooms, and that the student body (by grade or subject area) is tested at regular intervals. Classroom measures are applied to determine the effectiveness of the specific curricula and instructional methodologies, while school-wide measures are utilized to assess the comparative effectiveness of the general instructional program of the institution. All too often, direct observation indicates that there is an unfortunate disparity between accepted measurement techniques based on sound theory proven by research, and the evaluation procedures which take place in the classroom and even in school-wide testing situations (O'Donnell, 1981; Popham, 1981). It is the responsibility of the principal in the common school situation to ensure that proper student

measures are being employed at the classroom level, school-wide measures are coordinated, and that their results are properly utilized. Unless these principals are sufficiently familiar with state-of-the-art knowledge of student monitoring techniques, there is little chance that this responsibility can be satisfactorily fulfilled.

The problem can be more specifically defined by considering the following questions:

1. What is the a priori level of knowledge on the part of the principals, with respect to the effective production and use of student achievement measures?

2. Are differences in this level of knowledge associated with certain demographic information such as years of experience, grade level of position (elementary or secondary), and the size of the school?

3. Can a one-day, skill-building activity significantly elevate the principals' knowledge levels in the area of student achievement measures?

4. What is the effect of giving a group of subjects a pre-contact package of material containing an outline of the content of the presentation, as well as selected readings on the topic?

Purpose of the Study

Since principals are charged with the responsibility of maintaining a systematic student monitoring program, it is up to them to acquire the expertise necessary to accomplish this task. Given the limited resources of the school district and the restrictive demands on the principals' time, it is the purpose of this study to:

1) determine the present level of knowledge on the part of the typical principal and teacher with respect to student achievement measures;

2) determine the effectiveness of a one-day, skill-building activity designed to elevate principals' and teachers' knowledge levels concerning the production and use of student achievement measures;

3) provide the principals with a methodology which they may employ in helping teachers under their supervision to make more effective student achievement measures, and utilize the results of those measures in a productive manner; and

4) determine whether the results of a one-day training session can be improved by providing background material (presentation outline, handbook, and glossary of terms) to the target audience approximately two weeks prior to the training session.

Objectives of the Study

In order to accomplish the goals of this study, it will be necessary to:

1) develop an instrument designed to measure cognitive skills in the area of student achievement measures to be used in pretesting and posttesting workshop participants;

2) develop a skill-building program to deliver to principals dealing with the topic of student achievement measures;

3) produce an informational packet to send to a selected group of subjects taking the pretest approximately two weeks prior to the large-group instruction skill-building program;

4) deliver a skill-building program to principals and teachers, and collect pretest and posttest data from the group; and

5) analyze the data collected and refine the program accordingly.

Hypotheses to be Tested

The primary question addressed by this study is: Can a one-day, large-group instruction tutorial effectively raise the knowledge levels of principals in the area of student achievement measures? This question can be stated in the form of an hypothesis: Posttest scores will be significantly higher than the pretest scores which are measured for a group of subjects receiving a large-group instruction tutorial dealing with student achievement measures.

The primary question actually addresses two issues. First, what is the value of having principals well-trained in the area of student achievement measures? Second, will a program of this type provide significant training in this area? The first issue refers to the fact that principals need a sound background in the production and use of student achievement measures. This is necessitated by the fact that the principals supervise teachers' use of such measures and act as a resource person when professional development is needed in this area. The second issue refers to the tutorial presentation. Most principals have had classroom experience and, therefore, have been exposed to some form of training in the area of student achievement measures. However, this experience would not have included a methodology for supervising other professionals and subsequently instructing them in the use of such measures. The true value of the tutorial presentation is that it will act

to inform the principals as to the present state-of-the-art technology in the area of achievement measures. Secondly, it will provide them with a valid and functional program which they may institute immediately with their teachers to improve the classroom and school-wide use of student achievement measures.

Questions of secondary importance which will be addressed by this study are:

1. Are there any significant differences in pretest scores of subjects grouped by selected demographic data?

2. Are there any significant differences in posttest scores of subjects grouped by selective demographic data?

3. Are there any significant differences in posttest scores between the group which received the pretest and content information beforehand as compared to the group which took the pretest on-site?

Basic Assumptions

1. Subjects will be receptive to the skill-building experience and participate accordingly.

2. The subjects' responses to the pre/posttest instruments represent their true cognitive level with reference to the material tested at the time of testing.

3. Raising the subjects' level of knowledge about student achievement measures will assist them in helping teachers under their supervision make better use of same.

4. Research and current literature has identified effective techniques for the preparation and use of results obtained from student

achievement measures (the researcher will not be field-testing the strategies).

5. The group who receives the pretest in advance along with content information will not share this information with others who may be at the skill-building tutorial and be taking the pretest on-site.

Delimitations or Scope of the Study

1. While the researcher will be presenting background information on both standardized norm-referenced tests and domain-based criterion-referenced tests, an emphasis will be placed on the criterion referenced measures.

2. The principals to be sampled will largely come from groups seeking training on the subject material offered. As such, they may demonstrate a more positive disposition toward the type of program presented than subjects chosen at random.

3. Although accepted elements of learning theory indicate that it is preferable to "stretch out" the instruction over several sessions (in order to have periodic reviews), the tutorial will be limited to a one-time, large-group instructional experience. This is necessary due to the constraints on the accessibility of the subjects and other resource limitations.

Human Subjects Release

The Iowa State University Committee on the Use of Human Subjects in Research reviewed this project and concluded that the rights and welfare of the human subjects were adequately protected, that risks were

outweighed by the potential benefits and expected value of the knowledge sought, that confidentiality of data was assured, and that informed consent was obtained by appropriate procedures.

CHAPTER II. REVIEW OF THE LITERATURE

Introduction

Developments in the field of education have brought about important changes in the theory and the practice of both classroom instruction and student achievement testing. One theme common to these developments is the insistence that there should be a clear relationship among the instructional objectives, the classroom instruction, and classroom testing strategies employed by the teacher (Shoemaker, 1975). "Any form of systematic instruction has three common elements: (a) statements describing the intent of instruction, (b) instruction that is designed to help the student achieve the intended outcomes of the instruction, and (c) criterion-referenced tests that are explicitly related to both intent of instruction and instruction itself" (Roid and Haladyna, 1982).

Students preparing to enter the teaching profession are exposed to numerous classroom hours devoted to training in writing instructional objectives in terms of student behavioral outcomes. These same students are schooled in varying methodologies or models of classroom instruction. Common wisdom holds that there is an important link between instructional objectives and the teaching episode. This link must be made early in the training of every novice teacher. How well a teacher teaches to planned objectives can be determined by the appropriate use of classroom measures. The unsettling fact, however, is that little or no time is spent in guiding prospective teachers in the development and use of evaluation instruments for their own specific needs (Marshall and Hales, 1971; Green, 1963). When teachers are questioned about their background in the area of

student assessment, it is found that few teachers have had the chance to take courses in student assessment strategies. Those who have had such courses often feel that the content failed to deal with the teachers' immediate and pressing everyday assessment needs (Goslin, 1967; Captrends, 1984). Clearly, there is a felt need in the teaching profession to educate prospective teachers in the construction and use of student achievement measures (Ahmann and Glock, 1981).

Once educated in the procedures of classroom testing, teachers must be given assistance by their respective supervisors (or principals) so they may effectively integrate these procedures into their instructional repertoire. Learning to perform a new skill or strategy is only the first step toward effecting student outcomes. "Transfer of training to the learning environment requires skillful decision making by the classroom teacher and redirection of behavior until the new skill is operating comfortably within the flow of activities in the classroom" (Showers, 1982). An effective and systematic methodology for assisting the accommodation of new skills into the teacher's instructional repertoire is called the "coaching of teachers," and is described by Joyce and Showers (1982) and Fornies (1978). In order for this literature review to be meaningful, it must deal with two seemingly separate topics; specifically, classroom testing techniques and the coaching of teachers by principals.

The Coaching of Teachers

The most traditional approaches to staff development include such strategies as after-school inservice classes, "outside expert" lectures, university courses, participation at conferences, professional reading,

and possibly extended leaves (sabbaticals) for the purpose of study (Lieberman and Miller, 1979; Steig and Fredrick, 1969). Although these strategies are not without merit, they may tend to fall short of fulfilling their intended purposes (Rebore, 1982; Berliner, 1982; Houston and Freeberg, 1979). Simply put, the reason for any staff development activity is to change the behavior of teachers in a manner which will ultimately benefit the students. However, exposure to new techniques, no matter how worthwhile, does not insure the effective utilization of same in the classroom setting (Joyce and Showers, 1982). Through direct and personal contact with an individual acting in the coaching capacity, newly learned skills can be applied, practiced, modified, and ultimately integrated into the teacher's specific classroom situation.

The first step in any systematic professional development program is to determine the specific needs of the staff (Rebore, 1982; Friedman et al., 1980). Traditionally, this is accomplished by some sort of staff-wide needs assessment program which may take the form of a survey, meeting of the respective supervisors or principals, analysis of staff observation instruments, commercially available needs assessment kits (such as the one supplied by Phi Delta Kappan, the Flanders Interaction Analysis System, or the EPIC system) or examination of student achievement scores. (Occasionally, needs will be externally placed on the organization such as those indicated by legal mandate, contractual obligations, or community requirements.) Unfortunately, these classical types of macroscopic needs assessment strategies usually result in a variety of inservicing activities which tend to broadcast spray the staff

with a wide spectrum of information. As far back as 1935, James B. Conant opposed these practices of indiscriminate inservicing for teachers, although these methodologies are considered by some to be the most cost effective in terms of the number of staff contacted per dollar (Morant, 1981). However, it is of little wonder that some of the most common complaints of the staff receiving this type of treatment is that it is too general, theoretical, and simply does not meet their needs (Bierley and Berliner, 1982; Houston and Freeberg, 1979; Joyce and Showers, 1981). Staff members finding themselves frequently subjected to this type of collective inservicing quickly become little more than passive recipients of a mass of general information.

Contrasted with this seemingly crude but widely employed strategy, is the concept of coaching teachers for improved performance. This may be accomplished on an individual basis or in small groups. There are four steps which usually precede the actual coaching process. (1) As before, the needs of the teacher are first determined. Unlike the traditional holistic staff inservice attempts, however, this procedure focuses on the needs of the individual teacher. Among the ways that needs can be determined are direct observation, artifact collection, analysis of student scores, and possibly the request of the teacher involved. This assessment is also influenced by the teacher's content area, career stage, and specific needs of the students in the class. (2) The next step is to search the literature and research base in an effort to find a theoretical footing or rationale for an alternate teaching method which will fill the established need. To illustrate the procedure thus far, let it be assumed

that as a result of a direct classroom observation conducted by a supervising principal, it is determined that teacher X plans lessons well. Unfortunately, he/she consumes an inordinate amount of classroom time getting the students on task at the onset of the lesson. During the post-observation contact with the principal, the discussion reveals that this is not an uncommon problem, and both teacher X and the principal agree that it is important to remedy the time-wasting situation. Even a cursory review of the research on productive teaching models reveals several which directly address teacher X's dilemma. Rosenshine's Direct Teaching Model and Madeline Hunter's Seven Steps in the Teaching Episode specifically recognize that techniques such as review and preview, the use of structuring comments, and the establishment of anticipatory mind set on the part of the students help to get the lesson started promptly and assume proper direction (Rosenshine, 1980; Hunter, 1976). In this and many other cases, the teacher will have little trouble locating sound theoretical foundations for a solution to the specific problem at hand.

The teacher then (3) observes this procedure as it is demonstrated by someone who is relatively expert in the use of the desired teaching model. If the model is currently being utilized by another staff member in the school organization, there will be little trouble with this step. If this is not the case, alternate sources for such observations are other schools, contact with field experts, or professionally prepared video media (available from educational organizations such as the Association of Supervisors and Curriculum Directors or researchers such as Manatt at Iowa State University). (4) Practice and feedback comprise the next step in

the acquisition of new teaching skills. These are done under "protected conditions," such as with other faculty or with students who are characteristically easier to teach. Feedback may be provided through direct observation by another faculty member, by the supervisor, or by video taping the lesson and replaying it for the teacher or other selected observers.

There is nothing novel about the cycle of demonstration, practice, and feedback utilized to master a new skill (Chandler, 1978). These first steps which precede the coaching phase are specifically directed toward the development of a new skill on the part of the teacher. The final step involves the transfer of that technique into the classroom situation. A review of the literature from the field of psychology reveals that the term transfer refers to "the influence of prior learning upon later learning" (Klausmeier and Davis, 1969). There is a distinction made between lateral and vertical transfer (Joyce and Showers, 1981). Lateral transfer takes place when a person generalizes learning to a new task which exhibits a similar degree of complexity. One illustration of this is the case where a teacher successfully utilizes probing techniques in a mathematics classroom and decides to employ this same technique with a science class. The degree of complexity of subject material is very similar between the two class situations; therefore, the lateral transfer of this teaching technique should proceed smoothly. Transfer is considered vertical when abilities demonstrated in performing one task facilitates the performance of higher-order tasks. This is in evidence when a teacher is taught how to use a new methodology for classifying and

filing good test items for future use, and then modifies the technique to fit his/her own classroom teaching situation. In essence, the person has elevated the skills acquired as a student to his/her teaching situation. The differentiation here is due to the fact that the skills were acquired by the person in the student's role. These skills were then utilized in the much more complex role of the classroom teacher. The ultimate goal of the coaching process is to facilitate the cumulative transfer of skills and techniques learned in a training environment to the teacher's specific classroom situation.

The literature of supervision reveals several models which are similar and appropriate to follow in the coaching of teachers (Fornies, 1978; Friedman et al., 1980; Joyce and Showers, 1982; Bierley and Berliner, 1982; Ellis, 1965; Gross et al., 1971). The model offered by Joyce and Showers is one of the more recent in the literature and embodies most of the major elements of the others. It is for this reason that it will be the one used most frequently throughout this review of the literature.

According to Joyce and Showers, the practice of coaching involves five major functions:

1) provision of companionship,
2) giving of technical feedback,
3) analysis of application--extending executive control,
4) adaptation to the students, and
5) personal facilitation.

It must first be noted that sound pedagogical practice indicates that internalization of a new skill must begin right after training in order to be successful. As more time lapses between training and practice of new techniques, the chances that there will be a loss of subtle (but important) particulars is greatly enhanced.

With this fact in mind, the teacher must enter into a mutually advantageous partnership with another party during training or immediately thereafter. Ideally, this companion is another staff member seeking the same type of expertise. There may be occasions when a well-respected and trusted administrator may act in the desired capacity.

There is support for the view that teaching is a lonely profession (Rosenholtz and Kyle, 1984; Schwanke, 1982). Teachers feel isolated from their peers and in spite of the fact that they are in constant contact with large numbers of students, there are few opportunities for meaningful professional exchanges with fellow staff members. This type of support is fundamental to the successful integration of new skills on a classroom basis. Staff members should seek another professional with whom they can share frustrations, successes, and think through mutual perceptions and problems. Carl Rogers describes such a partnership as a "helping relationship." "This relationship is one in which at least one of the parties has the intent of promoting the growth, development, maturity, improved functioning, and improved coping with life of the other" (Rogers, 1961).

There is precedence in the field of education for such a relationship. In the states of New Jersey, Texas, and Michigan, the

helping teacher is a legally mandated position and an employee of the state who travels from school to school at the request of individual teachers (Lieberman and Miller, 1979). It is common in all states to provide a master or participating teacher as a guide for preservice education students. Studies have shown that at various times throughout the career stages of a teacher, he/she would benefit from the mentor/protege relationship. Under these conditions, both parties derive benefits from their mutual input and experiences.

Once the individuals have entered into this supportive relationship, they can provide the necessary technical feedback to each other as they practice their new instructional strategies. This technical feedback helps to ensure that growth extends to classroom practices. It has long been recognized by athletic coaches that the acquisition of new skills often has the immediate effect of diminishing the performance of the subject while he/she is attempting to cope with the formidable task of incorporating the new skill into a smoothly operating routine. "We'll generally make you worse before we can make you better" (Coach Richard Brooks of the University of Oregon to his incoming freshman football players ("Brooks Eyes New Season," 1981)). A strong helping relationship with provision for constructive technical feedback will help the teacher to overcome the frustration and feeling of failure which so often accompanies the utilization of a novel instructional strategy.

Analysis of application refers to the decisions made by the teacher with reference to the appropriate timing and utilization of new teaching skills. Examination of the curriculum and teachers' long-range plans will

reveal clues indicating when and where the newly acquired skills can be most advantageously employed.

It seems almost trite to mention that certain instructional techniques are more appropriate for specific types of students, but considerable research has been dedicated to fitting teaching styles and strategies to students of particular descriptions (Roscoe and Peterson, 1982; Stensrood and Stensrood, 1983; Willett, 1983; Goodlad, 1983). The fact remains that successful teaching is measured by successful learning, and anyone actively involved in the field of education will attest to the fact that students exhibit a wide spectrum of learning styles. Whether or not a new teaching strategy is successful is not determined by the degree of comfort or acceptance by the teacher, it is dependent upon how well the students respond to the new treatment. To determine this element, the teacher must be ready to employ generous amounts of formative evaluative techniques (Scriven, 1967), such as frequent quizzing (graded and ungraded) and probing questions. This must be followed by a comprehensive summative evaluation with high content validity (Popham, 1978; Ahmann and Glock, 1981; Anastasi, 1976; Ebel, 1965) to check for student mastery of material presented (Bloom, 1976; Green, 1963).

Personal facilitation is the final step in the coaching process. The successful use of a new teaching method or tool requires practice. Even after the first four steps in the coaching process are successfully accomplished, the newly acquired skill becomes perfected and can be used to its maximum potential only after it has been practiced and "overlearned." Overlearning refers to the state achieved when a skill is

utilized so smoothly and effortlessly that its use becomes second nature to the teacher. The coaching companion's role during this stage is one of occasional observer and confidant. It must be kept in mind that in a classroom situation where a teacher has executive control over the proceedings during the course of the day, it is ironic that the teacher is invisible only to him/herself. The value of an experienced observer acting as a mirror on the classroom has the potential for providing valuable information on the progress of the instructional strategy.

Coaching of teachers is a skill well worth mastering for the supervisor or principal. However, coaching assumes the existence of a close working relationship among the parties involved and should be employed where such a relationship is possible. Also, coaching is most effective when used to master a definable skill or demonstrable technique and is not intended as a cure-all to be administered on a broad scale in the hopes that it will cause the staff to get better at their jobs. The degree to which a single teacher will benefit from coaching is first dependent upon the ability of the supervisor to correctly diagnose the unsatisfactory behavior and find a desirable replacement behavior.

Classroom Testing Techniques

The coaching process takes place only after a new skill has been introduced, researched, and practiced under simulated or "safe" conditions. The principal can be instrumental in facilitating the coaching process by making the initial classroom observations which help to pinpoint areas of potential growth on the part of the teacher. The principal can also play a fundamental role in the process by introducing

the teacher to the new skills and strategies which will become the focus of the professional development experience. The following portion of the literature review presents a brief historical perspective on testing procedures as well as a look at the skills necessary for developing and using student achievement measures in the classroom.

There are 17 separate divisions in the field of applied psychology recognized by the American Psychological Society. One technique which is employed by all divisions to varying degrees is that of testing. The science of testing is commonly called psychometrics (Enc. Americana, vol. 22, 1960). Its goal is to investigate the development and administration of mental measures as well as to interpret the results of such instruments by the application of the appropriate statistical methods. In order to investigate the background of the psychometrics movement, it is first necessary to define the term most closely associated with it, namely testing. Although the definitions are as numerous as the authors who write them, there are a few which fairly represent their basic themes. Lee J. Cronbach defines testing as "...(a) systematic procedure for observing a person's behavior and describing it with the aid of a numerical scale or category system" (Cronbach, 1975). Lidz points out that testing is only "...a sample of behavior under controlled conditions" (Lidz, 1981). Clift and Imrie (1981) describe a test as a device which assesses the "...absolute performance with reference to a course of study." Considering these among other definitions leads the investigator to a clear understanding of the fact that the concept of testing involves at least three dimensions. First, testing deals with behaviors. "It is

important to point out that we never measure or evaluate people. We measure or evaluate characteristics or properties of people" (Ahrens and Lehmann, 1980). Second, testing is limited to a discreet and finite sampling of these behaviors. In point of fact, one measure of the quality of a test is how well it samples the universe of behaviors it purports to measure. Finally, by their very nature, tests result in some scoring strategy which allows the tester to make decisions about a specified performance of an individual taking the test (Ashworth, 1982; Burns, 1979). With these three aspects of testing in mind, it is now possible to study the roots of the psychometrics movement.

Most historical perspectives dealing with the formal use of measuring devices (tests) begin with some mention of the Chinese civil service tests during the Chou period 1027-256 B.C. (Enc. Americana, vol. 7., p. 28, 1960). The rigorous examinations not only served to discriminate among the test takers, but also produced a stabilizing effect on the nation (Smith and Adams, 1972). By utilizing a standard language, maintaining ancient customs and traditions and essentially dictating the curriculum which would govern societal life, these tests accurately reflected the goals and objectives of the culture at that time.

Perhaps the ultimate use of testing for purposes of discrimination among individuals was the one administered to the Ephraimites by the Gileadites (Gibeonites) in the ninth century B.C. (Judges 12:4-6). The Gileadites held the ford over the River Jordan, and, in order to prevent the Ephraimites from crossing, a peculiar but effective test was employed. The criterion for passing the examination was the correct pronunciation of

the word SHIBBOLETH. The Ephraims could not correctly pronounce the "sh" blend. Passing the examination allowed one to proceed unmolested, while failure resulted in summary execution (Enc. Americana, vol. 12, p. 653, 1960).

During the fifth century B.C., Socrates developed a style of formal, but unwritten, instruction and testing employing the use of both direct and leading questions (Green, 1970). These tests were asked of an individual in front of small groups, and the answers were given verbally to the group as a whole. This technique and its variations are still employed and are called Socratic questioning, after its author.

With the coming of the Christian era, various organized religions have employed the catechetical method of teaching and testing. (Augustine's treatise, "First Catechetical Instruction," appeared circa 405 A.D.) While this system is still in use today, it may represent the first example of a published and standardized set of questions and responses used to measure knowledge gained as a result of instruction.

According to authors Smith and Adams (1972), one of the first formalized, written examinations used for admission as well as for purposes of determining mastery, occurred at the British University in the 1760s. These tests were given at the end of each term and determined eligibility for advancement to the next level of instruction. American universities began employing similar testing strategies during the 1830s.

In 1816, Fredrich W. Bessel, a German astronomer and mathematician working at the Greenwich observatory, postulated that phenomena such as human behavioral differences could be subjected to numerical description

(Green, 1970). Bessel's conclusion precipitated from a review of a 1796 report which described the dismissal of an observatory employee. It was recorded that this worker's reaction time was characteristically 1.5 seconds slower than what was considered normal for recording the occurrence of celestial events. The amazing consistency of the worker's incorrect readings captured the attention of Bessel and caused him to recommend that statistical analysis techniques are appropriate for interpreting human behavioral characteristics. This conclusion became part of the foundation upon which modern normative-referenced evaluation schema are supported (Tuckman, 1975).

In 1863, Sir Francis Galton, a half-cousin to Charles Darwin, began his systematic study of human anatomical and behavioral differences. While most of his measurements were of basic human functions and reactions such as seeing, hearing, quickness of blow, sensitivity to pain, Galton searched in vain for a connection between these physical characteristics and mental ability. Although unsuccessful in his efforts to correlate physical and mental ability, Galton's work yielded significant statistical techniques and spawned a number of other studies investigating the uses of mental testing (Enc. Americana, vol. 15, p. 67, 1960; Tuckman, 1975; Green, 1970).

Prior to the 1840s, the Boston Public Schools had utilized a formal but unwritten strategy for student evaluation. Students were allowed to pass on the basis of oral examinations, speeches, and in some areas, proof of technical expertise by observation of specified performances. About 1845, the enrollment of the district began to grow to the extent these

assessment practices became unwieldy and inconsistent in their ability to discriminate among students. As a result of this fact, and influenced by the strong direction of its first State Secretary of the Board of Education, Horace Mann, the school district moved toward a system of formal, written examinations. An interesting characteristic of these examinations is that they were administered to students with no advanced warning (Smith and Adams, 1972). In fact, the content of the test and the proposed time of administration was a closely guarded secret within the faculty body. Mann listed the following advantages of written tests over unwritten examinations: written tests are impartial, fair, thorough, and they prevent interference by the teacher and favoritism while making the results known to all. One other subtle point Mann delineated was that written tests allow for the examination of the questions in the test, as well as the student taking the test (Smith and Adams, 1972).

In the 1890s, an American physician by the name of Joseph M. Rice developed what can be considered to be the first standardized achievement test in spelling (Green, 1970; Smith and Adams, 1972; Popham, 1981). This was quickly followed by standardized achievement tests in math and language usage in 1902 and 1903, respectively. It must be noted at this time that the term "standardized" refers to the test form, conditions for test administration, and to scoring procedures (Ahmann and Glock, 1981; Popham, 1981; Mason and Bramble, 1978). In the case of Dr. Rice's tests, the form was written response, the conditions were such that the questions were orally administered to groups of students, it was a timed test, and a formal scoring key was issued to the respective scorer. The term

standardized did not then, nor does it now, imply whether the test scores are computed using a standard normal curve as a reference, or whether a specific minimum number of correct responses established the criterion for passing.

Until the early twentieth century, classroom testing generally took the form of written responses to items read aloud in front of groups of students. These conditions were necessitated by the absence of duplicating equipment and the paucity of consumable resources such as paper. The next step in the evolution of formal classroom measures was the utilization of samplers or workbooks, once again, nonconsumable. The students would be given the written test forms at the beginning of the period, and the booklets would be returned unmarked at the end of the testing period. Responses were written on paper supplied either by the student or the institution.

Around the turn of the century, essay tests were beginning to find popular use (Lindquist, 1951). They generally consisted of characteristically brief questions which initiated rather lengthy answers. Emphasis was placed on how the student defended his or her assertions, rather than on the cold accuracy of the responses. It was thought that the use of this testing device represented the optimal advantages of the written examination (as outlined by Mann), while demanding that students be able to express their thoughts logically and in a coherent fashion.

About this same time, an educational philosopher and scholar by the name of E. L. (Edward Lee) Thorndike (1874-1949) began work on what would become a series of definitive documents and test samples dealing with a

wide spectrum of topics ranging from arithmetic (Stone Arithmetic Test) to handwriting (Thorndike Handwriting Scale) (Popham, 1981; Smith and Adams, 1972). The tests Thorndike produced were standardized, written, group achievement examinations and were accompanied by detailed documentation explaining the development of the questions as well as the use of the instrument. Oddly enough, Thorndike's original motive for developing formal achievement tests was not to assess student progress or to improve teaching, but rather to establish the profession of psychology as a science separate from philosophy (McKenna, 1977). It was Thorndike's contention that human behavioral testing was a tool employed by psychologists, not philosophers. Thorndike is also known for his authorship of one of the first test and measurement handbooks entitled "Mental and Social Measurements" in the year 1912 (Ebel, 1965) and the development of what has come to be known as Thorndike's Law of Effect. This edict simply states that people tend to repeat those behaviors which result in satisfaction, and avoid those which result in dissatisfaction. From this assumption, Thorndike reasoned that success on tests, good grades, and praise could be used as positive reinforcers in a classroom situation. In a later work which actively supports this view (The Technology of Teaching, 1968), B. F. Skinner postulated that intrinsic factors such as sheer knowing are of secondary importance to students. Instead, tests become contrived reinforcers and provide visible and reliable short-term goals in a classroom setting. Likewise, other contemporary authors maintain that one viable role of tests in a classroom

is that of a goal-setting device and reinforcer (Green, 1963; Ashworth, 1982; Clift and Imrie, 1981).

The tale is well-known of how the French Minister of Public Instruction commissioned a physician and psychologist by the name of Alfred Binet to construct a test which could be used to identify students who displayed low scholastic potential. The ultimate purpose of this procedure was to identify those students who would benefit from the normal scholastic program then in use (Lindvall, 1967). In 1905, Binet and Theophile Simon published the first standardized aptitude scale (Popham, 1981; Borg and Gall, 1983). The test was individually administered, and the term "intelligence test" was strictly avoided by Binet since he was a firm believer in the fact that intelligence was not a fixed, measurable quantity (McKenna, 1977). It was not until 1912 that L. Wilhelm Stern, a German psychologist, introduced the concept of Mental Quotient (Burns, 1979). This MQ (as it was called at the time) was derived by dividing the accumulated score obtained on the testing device, referred to as the mental age, by the chronological age of the student. This concept is the forerunner of the modern ratio IQ, the only difference being that the modern IQ is equal to the MQ multiplied by 100. Since that time, both the MQ and the ratio IQ have been discontinued in favor of the standard deviation IQ. This IQ is a standardized score derived from an appropriate frequency distribution having a mean of 100 and a standard deviation of 16 (Durost and Prescott, 1962).

The idea of testing human potential was very attractive and received a lot of attention on both sides of the Atlantic Ocean. In 1916, Lewis

Terman of Stanford University developed the first battery of tests which were specifically designed to measure a person's intelligence quotient. Unlike Binet, Terman strongly believed that human intelligence was a quantifiable and (allowing for age), stable trait which could be attributed to each individual.

Since that time, the testing movement has burgeoned. The list of contributors to the modern psychological testing movement is legion, and many "firsts" are claimed depending upon the information source and the construct being measured. However, no list of contributors would be complete without mention of the following examples. The first group administered aptitude tests were produced by Arthur Otis in 1917, Free Association Tests by Hermann Rorschach in 1922, Two-factor Aptitude testing by Charles Spearman in 1904 and again in 1928, Interest Inventories by E. K. Strong in 1927, Thematic Apperception Test by Henry A. Murray in 1943, and the Adult Intelligence Scale by David Wechsler in 1955 (Green, 1970).

Concurrent with the advances in psychometric instruments, the eight-year study (1932-1940) was conducted under the direction of Ralph Tyler and others (Aiken, 1982). One of the school variables this study examined was the production and use of teacher-made achievement tests in the classroom. This study emphasized the importance of such testing devices and encouraged the teaching of test and measurement techniques in the normal schools of education.

Just prior to the entry of the United States into World War I, Robert Yerks and Arthur Otis were commissioned by the Department of Defense to

produce a battery of tests which could be administered to draftees. The purpose was to discriminate between those individuals who might be suited to fill administrative and tactical positions as opposed to direct-line combat duty (Borg and Gall, 1983; Green, 1970; Smith and Adams, 1972). These tests were known as the Army Alpha (group written) and Army Beta (individual, nonverbal) forms.

With the advent of World War II, multiple aptitude tests for specific job descriptions and classifications were developed and employed to make placement decisions. There was also large-scale use of diagnostic and readiness assessments instruments to determine point of entry for various types of training.

The next 40 years were largely devoted to the refinement of individual testing techniques and to development of large-scale evaluation instruments and procedures. In recent times, testing research has been launched less for its purely scientific benefits, and more for its profitability. In short, testing has become big business. Several major publishing houses have devoted substantial resources to the writing and necessary documentation of their own test forms. These instruments are utilized by school districts, and in some cases by entire states for such diverse purposes as the assessment of pupil readiness, achievement, aptitude, individual interests, and perceptions. There are even tests available which are said to determine a teacher's qualifications for professional employment.

The federal government is still deeply interested in aptitude testing. With the growing use of high technology-oriented equipment,

there is a real need to place the limited numbers of inductees where their efforts will be most effective. This is especially true in light of the recent move to the formation of a volunteer army.

The next portion of this literature review is devoted to the discussion of test types and their uses in a contemporary context. Generally speaking, modern psychological tests may be classified in a number of ways. Tests may be categorized by their:

1) subject matter (spelling, arithmetic);
2) intent (discrimination power, ability to assess present or future status);
3) specificity of administration (standard, nonstandard);
4) the source of material included in the test (domain-referenced, general);
5) method of scoring (norm-referenced, criterion-referenced);
6) mode of administration (verbal, nonverbal);
7) response (restricted, free);
8) origin (teacher-made, published);
9) construct measured (achievement, aptitude);
10) level of difficulty (speed, power, mastery); or
11) response classification (cognitive, noncognitive).

This list is by no means exhaustive, since new tests are being prepared periodically (Ahmann and Glock, 1981; Burns, 1979; Cronbach, 1970).

While it is clear that analysis of a test could be made with reference to any of the classification schemes noted, this study will

begin by organizing tests into two basic categories--aptitude and achievement.

Aptitude tests

Aptitude may be defined as a talent, inclination, or tendency (Webster's Seventh New Collegiate Dictionary). In the field of psychometry, the term aptitude testing is used synonymously with intelligence or IQ testing and is usually used to determine general competencies or abilities (Popham, 1981). Interestingly enough, other authors contend that all tests measure learning. If this assumption is correct, the true distinction between aptitude and achievement tests is the fact that achievement tests measure specific learning acquired under relatively known and controlled circumstances. Conversely, aptitude tests measure generalized learning acquired under relatively unknown and uncontrolled circumstances (Ahmann and Glock, 1981; Anderson, 1981). Common examples of available aptitude of test are the Weschsler Preschool and Primary Scale of Intelligence (WPPSI), Weschsler Intelligence Scale for Children (WISC), Weschsler Adult Intelligence Test (WAIS), Peabody Picture Vocabulary Test (PPVT), the Stanford-Binet Intelligence Scale, the Scholastic Aptitude Test (SAT), and the Otis-Lennon Mental Ability Test.

By its very nature, aptitude, or general intelligence, is a comparative term and, as a result, these tests tend to be published so they can be widely distributed. These tests are also standardized, so the scores obtained are equally affected by test bias and norm-referenced so the scores obtained may be compared to other students' scores who took the exam (Hudgens and Phye, 1983; Clift and Imrie, 1981; Hively, 1974; Durost

and Prescott, 1962). In a general aptitude test, questions usually are intended to assess inductive and deductive reasoning skills; however, specific subject material may be drawn from virtually any of the curricular fields in education.

Authors jealously guard the specific item forms[1] for the test. In this way, it is hoped that the effects of coaching students for the test will be reduced, and the consequences produced by differential learning experiences of the test takers will be minimized (Anderson, 1981; Hively, 1974). In fact, aptitude tests are less concerned with what the student has already learned than his/her projected ability to learn new material.

The results of aptitude tests are generally used to classify students, predict their future success, select persons for fixed quota requirements, and to allocate limited resources (McKenna, 1977; Ashworth, 1982; Popham, 1981; Hively, 1974; Clift and Imrie, 1981). As Goslin states:

> It is impractical and undesirable to train every person for every position and have everyone try out for every position. One alternative is to use performance in school as an indication of general ability and to allocate opportunities for further training and positions on the basis of school performance. But, grading may be subjective and standards are different from school to school. Standardized (norm-referenced, aptitude) tests are the viable alternative (Goslin, 1967).

---

[1]An item form is the blueprint used to construct the items in a test. It contains information regarding how the questions will be asked and the expected response.

Oddly enough, in a later publication, Goslin contends that some cultures find this practice as anathema to their basic ideology. For example, the basic political beliefs of the Soviet Union preclude the assumption that differences in abilities among normal individuals are an inherited trait. Consequently, assessment has tended to focus around achievement differences among the perspectives. These differences in achievement are assumed to be based on motivation and individual effort, rather than on innate variability. It is for this reason that the Soviet Union institutes a state policy which works to motivate citizens, rather than to select recipients of limited resources on the basis of aptitude measures (Goslin, 1967). Unfortunately, there is often a disparity between policy and practice.

The successful aptitude test attempts to effectively discriminate among the test takers (Hively, 1974; Burns, 1979; Baglin, 1981; Durost and Prescott, 1962). Since the test results are used to make decisions about the comparative ability or potential of each student, the tests are designed to spread the scores obtained as much as practically possible (Ahrens and Lehmann, 1980). With this thought in mind, generally the difficulty index (sometimes referred to as the facility index) for the items chosen for the test is kept at or about the .5 level (Smith and Adams, 1972; Hudgens and Phye, 1983; Clift and Imrie, 1981). This means that each question (or parallel forms of each question) must be pretested

with representative groups, and that ideally 50 percent of the students attempting that item answered it successfully.[1]

Another important fact about aptitude tests is that aptitude measures are designed in such a way that virtually no students will obtain a perfect score (Hively, 1974). This occurrence would prohibit successful sorting among the students having the best scores. Likewise, as a consequence of the normalized ranking strategy, the average score obtained by the test takers falls at a percentile rank of 50 percent (McKenna, 1977).

Aptitude tests can either be formed nationally or regionally. A nationally normed test includes scores from representative samples of students tested across the entire country. A regionally normed test may base the scoring and ranking procedures on certain regions such as the northeast or western U.S. It is also possible to obtain normative scoring information for selected large population areas such as specific cities or states. An interesting side note with reference to the normative scoring techniques is that at least one author questions the methodology which test publishers employ to arrive at their standard normal scoring curves. Baglin contends that student scores which are utilized in the statistical process of producing a standard scoring curve are self-selecting and do not necessarily represent a random sampling of all possible cases

---

[1]Difficulty index is computed as follows: D = number of students correct divided by number attempting the item. Ahmann and Glock point out that a difficulty index of approximately .5 is most appropriate for tests having low inter-item correlation, which is typical of aptitude tests.

(Baglin, 1981). This self-selection process arises from the fact that cases are obviously drawn from participating districts. These districts almost invariably employ textbooks prepared by the same company which publishes the aptitude test. Likewise, these districts tend to administer standardized tests from the same publisher (or parent company) on a fairly regular basis. Baglin claims that these practices make the staff and student body in these participating schools test-wise, and, therefore, there exists the potential of receiving artificially inflated grades. These grades are the sole basis for the production of the publisher's standard normal curve, and, therefore, a school which does not fit the description of the "average participating district" will run the risk of finding its scores atypical and possibly lower than expected.

The validity (or truthfulness--Green, 1963) of a test refers to how well the test measures what it purports to measure (Ebel, 1965; Mason and Bramble, 1978; Ahmann and Glock, 1981). Since the focus of a norm-referenced aptitude test is its success in differentiating among people, it is not necessary for it to be very specific about the subject matter it covers (Hively, 1974). What is important for a test whose planned use is to make judgments about the potential of a person is predictive validity. For instance, the Scholastic Aptitude Test, or SAT, is widely employed to predict future success in college. One measure of its validity is how well scores on the SAT correlate with cumulative averages of those same students in college. This is accomplished through follow-up studies conducted by school districts, universities, and by the test publishers.

Construct validity is also of interest in certain aptitude tests (Ebel, 1965; Ahmann and Glock, 1981). This is considered in two different ways--convergent and discriminate validity (Mason and Bramble, 1978). Convergent validity refers to how well the test compares to other accepted measures of the same construct. It is for this reason that test publishers are deeply interested in what other test publishers produce.

Discriminate validity refers to how well the test sorts students on the basis of the construct measured. This is usually established by a panel of experts (Borg and Gall, 1983; Mason and Bramble, 1978).

Reliability (or consistency--Green, 1963) of a test is usually defined in terms of the instrument's stability and can be interpreted as being either internal or external. Internal reliability refers to the extent of item homogeneity within a single test. This is determined statistically by any of the available reliability formulas, such as the Spearman-Brown 1/2 Test Formula, Kuder-Richardson KR20 or KR21, Hoyt's Anova Procedure, or Cronbach's Coefficient Alpha (Hinkle et al., 1979).

A test is considered reliable in terms of test/retest stability if the test tends to yield equivalent results over repeated administrations. This type of reliability is especially important in aptitude tests, since the value of such tests lies in the consistency of predictions made on the basis of the results obtained over time.

Achievement tests

Unlike aptitude tests, which attempt to make predictions about a student's future based on his/her scholastic potential, achievement tests measure what learning has occurred in the past (Anderson, 1981). As

mentioned previously, aptitude tests are necessarily published, standardized, norm-referenced, and contain a generalized subject content. Achievement tests, on the other hand, may or may not fit any of these descriptions. This portion of this review will first address standardized achievement tests. Examples of this type of test are the Woodcock Reading Mastery, Key Math Diagnostic Arithmetic Test, ACS-NSTA High School Chemistry Examination, and State Regents Examinations (by subject areas).

Material used for items on standardized tests is derived from specific written curricula such as the respective state Regents Syllabus, or from a generally accepted, although possibly unwritten, curriculum such as the one used to generate the California Achievement Test (CAT) (Zavarella, 1980). In either case, the results of such tests are utilized to determine the individual student's level of mastery of specific subject material. This is accomplished either with reference to some predetermined level of competency (as in criterion-referenced tests) or comparatively with reference to the rest of the population taking the test (as in norm-referenced tests) (Lidz, 1981; Thorndike and Hagen, 1977; Gronlund, 1972; Marshall and Hales, 1971). Clift and Imrie define level of mastery as "the proportion of material learned compared to how much he(/she) should or could have learned" (Clift and Imrie, 1968). Information gathered with respect to level of mastery is necessary for determining such things as preparedness for future learning, diagnosing student difficulties, certification for competency, eligibility for admission whenever a quota-free selection system is being used, and

finally, progress and growth in a subject area or discipline (Gearheart and Willenberg, 1974; Gorow, 1966; Edmonds (cited in Kelwin, 1984)).

Standardized tests may serve an important function apart from the direct assessment of students. These tests may be utilized by a school organization to determine schoolwide curriculum alignment (Zavarella, 1980; Cooley, 1980). Test publishers will readily provide documentation describing the domain from which the test content was derived. If the standardized test is drawn from a domain which the school feels is worthwhile, the results of such tests could indicate the degree of parallelism between the schools' curriculum and the domain of the test publisher (Hively, 1974).

Fenwick English describes a process for assuring classroom curriculum alignment as an exercise which first involves curriculum mapping (English, 1980). The purpose of this exercise is to identify precisely what has been taught in the classroom, not just the theoretical scope and sequence of the written curriculum. Once it has been established through this curriculum mapping exercise that the instructor actually taught those things in the written curriculum, the administration of standardized tests will aid in determining if the curriculum of the school is aligned with the domain of the test publisher.

In one study performed in the Pittsburgh Public Schools during the 1981-1982 school year, the utilization of a program of frequent student achievement monitoring appeared to be responsible for a significant level of curriculum alignment when compared to a control group using no such program (Le Mahiew, 1983). The results seemed to be linked to grade

level, since a stronger curriculum focus tended to occur at higher grade levels.

Standardized tests are also used for input in making personnel decisions. If there is appropriate correlation between the curriculum of the test and that of the school, the grades on tests might indicate the degree of success of the teaching strategy employed (Graeber, 1984). Consistently depressed test scores would indicate a close look at the present strategy. If a deficiency is diagnosed and a new strategy employed, the results of the intervention may be monitored by subsequent administration of parallel forms of the same test. This methodology has been employed successfully by programs such as the School Improvement Model (SIM) Project, Iowa State University (Manatt, Stow, and others). It must be noted that the success of this process is partially dependent upon assuring that the students admitted to the class have the requisite ability and motivation to achieve (English, 1980). These variables must be assessed before a realistic personnel decision can be made. As Lidz puts it, "The examiner must, of course, judge the appropriateness of the domain for the particular student. No student can be expected to be a master of all content areas" (Lidz, 1980).

With reference to the level of difficulty of standardized tests, there is considerable variation depending upon the purpose of the test. A mastery test would be employed to determine the progress of the student at a given point and would represent a moderate difficulty level. For diagnostic purposes, a power test might be used. In this case, questions at the beginning of the test have a low difficulty level, but it increases

throughout the examination. This allows the instructor to determine the specific areas or level of competence at which the student may need assistance (Ebel, 1965).

A speed test is constructed with a larger number of items than an average student is expected to be able to answer in a given period of time. The difficulty level is characteristically low for speed tests. The degree of success on this test is determined by how fast a student can work, and not by the difficulty of the tasks he/she can accomplish (Ebel, 1965).

Apart from standardized, published, achievement tests are teacher-made achievement tests. These are generally domain-based, criterion-referenced tests and are of particular interest because they are the measures most commonly applied in the classroom (Anderson, 1981). Although norm-referencing of teacher-made tests is possible, it must be done with great care. In order to utilize a standard normal curve, it must be assumed that the grades received on the measure are randomly distributed. This is seldom the case in a classroom situation, especially on the secondary school level. Students are usually grouped by ability, intentionally or otherwise, and, therefore, the raw scores tend to locate predominantly toward the top of the scale. Scoring in the standard normal curve could mean that missing one or two items on a 50-item test would cause a student's converted to standardized score to be significantly affected (Burns, 1979). Many times this difference in the raw scores (as represented by missing a very small number of items) falls well within the standard error of the testing instrument. This means that if the same

student took a parallel form of the same examination, he/she might happen (purely by chance) to get that same number of items correct and, therefore, receive a significantly higher score.

It is also of statistical importance that the number of cases included when generating the standard normal curve exceeds 100. In fact, 125 is the recommended minimum by many experts (Hinkle et al., 1979; Burns, 1979). Once again, this seldom happens in a public school classroom. It is much more common for class sizes to be less than 30. Including other classes in the scoring curve assumes that identical instructional experiences were received in all classrooms. Once again, this is rare in a classroom setting. Many shortcomings such as these few mentioned can be minimized by appropriate statistical methods, but these all require specific expertise and generous time requirements, neither of which are readily available to the average classroom teacher.

In any event, grading students on a standard normal curve results in a grade which does not truly reflect the degree to which the student has mastered the subject material (Popham, 1981; Zavarella, 1980; Clift and Imrie, 1981). It only indicates how well that student has fared when compared to the group he/she has been scored against (McKenna, 1977; Gronlund, 1972; Hively, 1974). The danger in this type of evaluation technique lies in the fact that the student's score depends not only on his/her ability in the subject area, but on his/her classmates' as well.

The most widely accepted strategy for avoiding the pitfalls of norm-referenced classroom grading is to grade a student against some predetermined level of competence as represented by a minimum passing

grade (Lidz, 1981; Martuza, 1977; Glaser and Nitko, 1971; Popham, 1978). A good definition of this type of criterion-referenced measurement is given by authors Smith and Adams: "A criterion-referenced test (CRT) is designed to determine whether or not a student has learned to perform a particular function at a specified level of quality" (Smith and Adams, 1972).

Criterion-referenced grading is most effective when applied to tests which are developed from a specified domain of information and behavioral objectives (Lidz, 1981; Borg and Gall, 1983; Harris and Stewart, 1971). This arises from the fact that mastery of the domain becomes the criterion for passing the examination. Since classroom instruction evolves from just such a domain, this measurement technique is most desirable for classroom use. CRT's are the means by which a classroom teacher can assess how well the student has mastered the subject material presented to him/her (Ahrens and Lehmann, 1980).

The level of difficulty varies with the use of the test (Ahmann and Glock, 1981). If the instrument is used as a pretesting device, the questions would be considered very difficult by students taking the test. In fact, few students, if any, should be able to answer even a small number of questions correctly. On the other hand, if the test is properly constructed to measure learning gained as a result of specific teaching and was employed as a posttest, the items on the test would be much easier for the students. It is hoped that every student will achieve to at least the minimum standard set for mastery.

The validity of teacher-made, criterion-referenced tests is established qualitatively and is dependent upon the instrument's alignment to the domain of presented material it hopes to measure. This alignment can be optimized by following a few logical steps:

1) Generate a table of specifications for each test.
2) Inform the students of the contents of the table of specifications at least one day preceding the examination.
3) Test often enough so that a test filling the average classroom period can contain enough items to properly cover the material.
4) Develop a test before the teaching takes place and use it as a guide for planning the instructional experiences for the students.

The methodology for generating a table of specifications is not difficult, and the time it consumes is well-repaid by the time it saves in the planning and instructional process (Shaw, 1977). Tables of specifications may take many forms, but a well-accepted format related specific subject matter to behavioral changes which the teacher hopes to initiate. It is usually represented as a two-dimensional grid with content areas listed along one axis and intended pupil behaviors listed along the other axis (Marshall and Hales, 1971; Ebel, 1965). Also included is an approximate teaching time allotment to each area and a specified number of questions which should be devoted to each portion of the topic (Ahmann and Glock, 1981; Hudgens and Phye, 1983; Green, 1963). This blueprint for teaching and testing may be saved and modified as

necessary to be used in future classes. An example of a table of specifications is included in Appendix D.

Informing the students of the contents of the table of specifications prior to the test gives the students a guide for studying and assures both the instructor and the students that all the material which is slated for inclusion on the test was actually covered during instruction (Hudgens and Phye, 1983).

One prerequisite for validity of a CRT is that the instrument effectively samples the domain from which it was drawn (Green, 1963). Considering the fact that classroom periods average 40 to 50 minutes in length, the number of test items is limited by this time constraint. In order that the test can include a representative sampling of each behavioral objective on the table of specifications, it is imperative that testing occurs for a relatively small number of objectives at one time. It then follows that testing must take place at frequent intervals during the instructional process.

Other pedagogically sound functions are served by frequent testing as well. Frequent testing causes frequent review of material on the part of the students. Learning theorists have long maintained, and subsequent research has confirmed the belief, that review enhances retention (Storey, 1970; Hudgens and Phye, 1983). The effect of frequent testing is to distribute practice over time, resulting in long-term retention. If testing occurs infrequently, this results in students cramming for the exam. This massed practice produces rapid learning, but poor long-term retention.

It also is well-accepted that what is tested is what is considered important and, therefore, that is what receives the most attention from the students (Ashworth, 1982; Clift and Imrie, 1981). Finally, the function of feedback to the students with reference to their progress allows the students to readjust his/her learning strategy on a continuing basis (Scriven, 1967; Edmonds (cited in NSPRA, 1981); Mortimere (cited in Rutter and Mortimer, 1979)).

A caveat with regard to periodic testing is voiced by at least one author. If testing becomes too routine and monotonous, it may cause the more able student to become complacent and retard his/her progress (Heywood, 1977). To avoid this situation, it is suggested that the type of test used in the classroom be varied from time to time. For example, if multiple choice items seem to be the routine, then short answer or some other type of free response item might be periodically employed.

One interesting and most valuable technique for the classroom instructor is to develop the testing instrument before the instructional experience has taken place (Hudgens and Phye, 1983). If this is done, the content of the testing instrument constitutes an integral portion of the framework for planning and organizing the teaching episodes.

Test-retest reliability is not a major issue with teacher-made tests, since it is likely that they may never be used again, at least in that specific form. It is questionable on sound pedagogical grounds as to whether tests should be reused, since it may have the effect of freezing teaching into a fixed pattern unresponsive to contemporary changes in the respective field (Smith and Adams, 1972). Most teacher-made tests are

designed for single use and may be filed intact for reference purposes. Although certain questions may be reused, it is not likely that the entire test will be regularly used again.

A quick check which a teacher may make with reference to the internal reliability of the instrument is to scan the corrected papers and note whether the historically better achievers and the historically lower achieving students have scored in approximately equivalent positions on the instrument in question (Durost and Prescott, 1962; Lidz, 1981). This is referred to as rank-order tendency.

One fact for the classroom instructor to keep in mind is that generally, a longer test is more reliable than a shorter test, ceteris paribus. Although a 100-item test is much less than twice as reliable than a 50-item test (Gorow, 1966).

The length of a normal teacher-made test may vary with intent, since a single-concept quiz will probably be much shorter than a full-unit test. In either case, the time allotment should be such that approximately 95 percent of the students have enough time to complete the test on time (Dunstan, 1969).

There are some problems which may arise as a result of the use of teacher-made measures. There is some speculation as to whether the majority of classroom teachers have the expertise to successfully design, produce, and administer valid and meaningful classroom measures (Zavarella, 1980). "There is a definite technique or method to good test-making. This technique requires practice and deserves a higher place

on the list of teacher training requirements than usually accorded" (Gearheart and Willenberg, 1974).

The other problems teachers face in the use of teacher-made tests include the heavy workload on the teacher, the potentially staggering recordkeeping involved, and the time taken from instruction to administer the test. Although none of these problems lend themselves to simple solutions, there are some techniques available to ease the burden imposed by the use of teacher-made tests. For instance, a technique which may effectively reduce the workload is the use and maintenance of a comprehensive test item filing system (Ahmann and Glock, 1981). These systems may be either handwritten (Mershon, 1982) or computerized (Baker, 1973). With the advent of the inexpensive personalized computer, comprehensive recordkeeping may be more easily accomplished with the aid of one of the many currently available teacher gradebook software programs available, such as Apple Grade Book (available from Career Aids, Inc., Nordhoff St., Chadsworth, California). With respect to the amount of time needed to test in the class period, it must be realized that the test becomes part of the learning experience by causing review, clarifying goals, and providing useful feedback to the students and teacher. As such, it deserves as much time as is prudent and appropriate use warrants.

Probably the most difficult philosophical decision a teacher must make with regard to the construction of a criterion-referenced measure is where to apply the cut-off point for the purpose of determination of mastery (Ebel, 1973). As authors Smith and Adams point out, "Grading criteria must be set high enough to challenge the best students, but low

enough to be a reasonable specification for the average student" (Smith and Adams, 1972). With this as a guide, the instructor must take into consideration the abilities of his/her students, difficulty of material, and the time allotment for the particular portion of the topic.

Summary

Classroom tests have evolved into much more than a ceremonial rite of passage for students. Tests are now used to explore learning deficiencies, motivate students, analyze curriculum, align teaching and learning styles, and their results significantly affect the future of both the student and the classroom teacher. The review of the literature reveals a need in the field of education to assure that teachers possess a firm background in the area of student achievement testing. The facts indicate that little is done during the teacher's own educational experiences to fulfill that need. Much of the training for this highly technical professional skill occurs in the school setting after the teacher is hired. Peers sharing their experience and teachers themselves finding what seems to work best apparently accounts for much of the professional development in the area of student achievement testing. In short, testing has been traditionally treated in much the same fashion as parenting. Just as it is assumed that age and experience breed the skills necessary for successful parenthood, it is assumed that previous educational experiences and a few years of employment will somehow cause the existence of those skills necessary for the production of a valid and reliable program for monitoring student achievement. In truth, although the testing movement has at times been marred by controversy or criticized

by the unfortunate misuse of individual instruments, the mysterious art of testing has matured into the precise science of psychometrics. The skills needed to apply this science are clearly defined and well-documented. It is now essential to transfer these skills into the classroom teacher's repertoire of professional activities.

While it is the task of universities to upgrade their instructional programs to better qualify new teachers in the systematic use of student testing procedures, it is the responsibility of individual school organizations to provide instruction as well as maintenance services to its existing professional staff in this area. Research indicates that one of the correlates among effective schools is strong leadership on the part of the principal. Consistent with this leadership role is the ability of the principal to act as a resource person and to take an active part in the professional development of his/her staff. Common wisdom dictates that a principal who is better informed is better able to inform others. What is then called for is a mechanism to instruct principals (and supervisors) in the proper methodology for the preparation and use of classroom achievement measures.

CHAPTER III. METHODS AND PROCEDURES

The central question addressed by this study was: Can a one-day, large-group instruction tutorial presentation effectively raise the knowledge levels of principals in the area of student achievement measures? In order to test this question, a one-group, pretest-posttest design was employed. A target group was selected, given a pretest to determine their a priori knowledge of the treatment material, a treatment program was presented, and a posttest was administered to determine the group's knowledge level after the application of the specified treatment. (Note Appendix A, pre/posttest document.) Also, certain demographic information was recorded at the time of testing for the purpose of determining whether there were any meaningful relationships between such variables as grade or position level, years of experience, position or educational background, and the test scores as disaggregated by such variables. (Note Appendix B, demographic questionnaire.)

Another issue addressed by this study was whether a group of individuals, who received background preparatory material prior to the treatment, would score significantly higher on the posttest instrument than a comparison group who received no such information. To investigate this issue, a sample was randomly selected from the target group and pretested off-site approximately two weeks prior to the application of the treatment. (Note Appendices C and D, pre-contact cover letter and test cover letter, respectively.) At the completion of this pretest, each participant was given a packet of background material in the form of a handbook containing an instructional outline, glossary of terms, and

instructional notes. This same handbook was presented to the comparison group at the treatment site after each individual handed in his/her pretest. (Note Appendices E and F, workshop participant handbook and handbook cover letter, respectively.) Each participant was then instructed to read the material prior to attending the large-group tutorial. This group, which will be referred to as the off-site group in this study, was then present for the same tutorial as the comparison group which was pretested on-site for the purpose of determining if the scores obtained from the two groups revealed a statistically significant difference on the posttest instrument.

In order to test the questions raised in this study, it was first necessary to develop and field-test a tutorial program created from the latest proven technology in the field of teacher-prepared student achievement testing. This phase of the study was accomplished at Iowa State University beginning in June of 1984. Two subsequent field tests of the presentation followed.

Field Test One, involving 67 subjects consisting of principals, supervisors, and teachers, took place at the campus of Iowa State University in June of 1985. The program was presented, feedback was obtained in the form of workshop analysis forms, and subsequent revisions were made based on that feedback.

Field Test Two--the revised presentation was given before 42 subjects in New York State at the Board of Cooperative Educational Services (BOCES District I) building in July of 1985. Once again, workshop analysis forms were utilized and the program was revised into its finished form.

The Subjects

The investigation focused mainly on principals and supervisors, while teachers were included for demographic comparison purposes. The target group consisted of 53 subjects from the Mason City, Iowa area. The presentation of the large-group instruction tutorial was held in March of 1986 and the data were collected at that time and subsequently analyzed. Approximately two weeks prior to the tutorial, a sample of 17 (one-third) of the larger group was randomly selected as members of the off-site group. At that time, the off-site group was pretested and given the handbook to study before the tutorial. Due to circumstances beyond the comparison of the study, a number of group members either missed the pretest or the posttest, thus rendering their scores unusable for purposes of comparison with the rest of the target group. The final number of usable scores in each group was 17 for the off-site group, and 27 were pretested on-site and used as the comparison group.

The Large-Group Tutorial

The treatment consisted of a large-group format tutorial. This design was chosen since it was felt to be the most efficient methodology for reaching the greatest number of subjects at one time. Although one of the goals of this study was to instruct principals in the state-of-the-art techniques of classroom achievement testing, it was clearly intended that those principals transmit that information to their respective staffs. It was further believed that the large-group format would allow the principals to transmit that information in a manner which was most time- and cost-efficient. The principals were given the same basic tutorial

which they might present to their teachers. They were also provided with materials which they might use directly with their teachers to integrate testing into the instructional planning process, as well as guidelines for writing test items and various other charts and tables which they can use as resource material for a staff development exercise.

The tutorial consisted of four modules as follows:

1. Module I - Classroom Achievement Tests; Foundations. This module builds the theoretical and practical framework for classroom testing techniques. This segment included general descriptions of tests, classifications, and uses of classroom achievement measures.

2. Module II - Item Types. This module clearly delineates the type of questions in use and specifies rules for writing each type. Included in this segment is a brief, self-help quiz on item types and their most appropriate applications.

3. Module III - Testing as a Planning Objective. This module demonstrates how the test document fits into the planning process and subsequently into the greater instructional picture. Specific behavioral objectives are linked to test items in tabular form on a table of specifications and production of the classroom test is integrated into the planning/instructional cycle.

4. Module IV - Making the Test. This final module illustrates proven procedures for producing valid classroom tests quickly and accurately. Item banking is explained and procedures for banking items manually, as well as with the use of computer software, are explained and demonstrated.

The entire presentation involved approximately five contact hours, exclusive of the time needed to pretest and posttest the group. One major advantage of separating the presentation into modules was that it could be presented in part to a group if so desired. Additionally, the tutorial could be presented over as many as four separate occasions which allows for greater scheduling flexibility for both the principals and the staff.

Collection of Data

Data collected during the study were in the form of pretest and posttest scores as follows:

1) pretest, off-site (treatment) group;
2) posttest, off-site (treatment) group;
3) pretest, on-site (comparison) group; and
4) posttest, on-site (comparison) group.

Both pretest and posttest answers were recorded by the subjects on optical scanning sheets (note Appendix G) and processed at the Iowa State University test scoring facility. The pretest/posttest document consisted of 50 objective items. Thirty items were of the multiple-choice variety and each contained a stem, one correct response and four foils. The remaining 20 items were of the true/false variety and each contained a stem, one correct response and one foil in the standard true/false format.

Objective items were chosen over other item types for three reasons. First, test and measurements experts maintain that the most appropriate item type for assessing in-depth cognitive behaviors relating to a highly specified domain is the objective item type (Ebel, 1965; Popham, 1981). Clearly, the domain containing information on teacher-made student

achievement measures is to be considered a highly-specified domain. Second, objective items are most appropriate when it is necessary to eliminate extraneous scoring variables such as spelling, handwriting, and language variations. Finally, objective items lend themselves better to standardized scoring procedures and accepted item analysis techniques.

In addition to test scores, the following demographic information was collected from each subject at the time of testing:

1) total years of experience in the profession;
2) educational background;
3) current position; and
4) level of assignment.

Anonymity of the subjects was preserved through the use of code numbers to identify each subject's test booklet. At the completion of the study, each school organization will receive a copy of pretest and posttest scores listed by identification numbers and an abstract describing the study and its results.

Additionally, there was an item analysis run for the 50 items on the pretest/posttest document (note Tables 1-4). It should be noted at this time that the pretest and the posttest instruments were identical. It was decided to use the same instrument in order to avoid parallel form variability in the pretest and posttest scores. Put simply, if two separate forms of the instrument were used for pretesting and posttesting, a portion of score differences before and after application of the treatment would have been expected due to the fact that the two instruments were different.

Statistical Analysis

The standard test item analysis performed at the Iowa State University test scoring facility provided the following information on the pretest/posttest instrument:

1. Kuder-Richardson 20 reliability estimate

$$r_{tt} = \frac{n}{n-1} \frac{s_t^2 - \Sigma qp}{s_t^2}$$

where q = portion of incorrect responses
p = portion of correct responses

2. Average test score

$$\bar{x} = \frac{\Sigma x}{n}$$

3. Standard error of measurement in raw scores

$$s_e = s_t\sqrt{1-r_{tt}}$$

4. Standard deviation

$$s_i = \sqrt{\frac{\Sigma(x_i-m_i)^2}{N}}$$

These analyses served as indicators of the overall characteristics of the whole test instrument. The scoring service provided information useful for the improvement of each test item as follows:

1) number attempting the item;

2) number omitting the item;

3) number answering correctly;

4) item difficulty (% responding correctly); and

5) item-score correlation--derived by point biserial correlation between item performance and total test score.

$$r_{pbis} = \frac{m_p - m_y}{s_y}\sqrt{\frac{p}{q}}$$

The design of the study determined the type of statistical analyses most appropriate for addressing the questions presented in Chapter I. The primary question on the study was: Can a one-day, large-group instruction format tutorial presentation effectively raise the cognitive levels of principals in the area of student achievement measures?

The methodology employed in the study to address this question was to compare mean test scores on the testing instrument before and after the application of the treatment. This was accomplished through student's t-test comparison procedures for paired samples. In effect, the tests determined the possibility that the gain scores achieved by the subjects could have been reasonably expected to have occurred by chance alone.

The secondary question in the study was: Are there any significant differences in post-treatment knowledge levels between a group which received the pretest and content information beforehand when compared with a group which took the pretest on site and received no content information prior to the workshop?

The methodology utilized to investigate this question was to first compare the pretest score means of the off-site and the comparison groups to determine the a priori equivalence of those two groups. If the pretest scores of the two groups show no significant difference, then there is a strong indication that the two groups are roughly equivalent in their knowledge of the subject material as measured by the test instrument. Conversely, if there is a difference between the two groups a priori, no

further investigation of posttest scores to determine significant differences would be defensible. Subsequently, a comparison was made between the posttest score means of the off-site and the control groups. This was an attempt to determine if the off-site group benefited from having received background information on the content of the tutorial, as reflected by higher posttest scores.

The third question raised by the study was: Are there any significant differences in post-treatment knowledge levels of the subjects if they are grouped by selected demographic information? This question was an attempt to determine if predictions could be made concerning either the a priori, or the post-treatment knowledge levels of subjects relative to specified demographic information.

To address this issue, the statistical analysis employed was a one-way analysis of variance (ANOVA) comparing pretest scores and demographic data and posttest scores and the same demographic data. This method is roughly parallel with the student's t-test, but it is more appropriate for use when comparing several factors at the same time. The following independent variables were tested:

1) position in the organization;
2) level (primary, secondary, central office);
3) educational background; and
4) years of experience in education.

The pretest and posttest scores were the dependent variables in this analysis.

CHAPTER IV. FINDINGS

Descriptive statistics relative to the pretest/posttest instrument are presented first. Following this section, descriptive statistics relative to the performance of the subjects on the pretest/posttest instrument as well as the investigation of associations among the demographic data and the achievement scores are reported. The remainder of this chapter will be dedicated to the consideration of each hypothesis.

Pretest/Posttest Instrument

The item analyses for the four test administrations are indicated separately in Tables 1 through 4. Examination of Tables 1 through 4 indicates that the average difficulty of the items used in the tests ranged from 44.8 to 67.5. It must be noted that different administrations of the same test instrument will invariably yield some dissimilar item analysis results, even though the test items are identical among all the instruments. This is naturally due to variations in respondents' answers, and not in the test items themselves.

It would be expected that the average difficulty index in the pretest would be somewhat lower than the average difficulty index on the posttest due to the subjects' limited knowledge of specific content material. Indeed, this is the case, since the average pretest difficulty indices are 44.8 and 49.8, whereas the average posttest difficulty indices are 61.2 and 67.5, respectively. The general indication here is that the test was measurably easier after the subjects attended the large-group tutorial.

Table 1. Summary of item analysis results for pretest, off-site (treatment) group

Item number	Discrimination	Difficulty
1	.48	47
2	-.21	24
3	.35	71
4	.00	100
5	-.14	18
6	.06	35
7	.67	12
8	.32	59
9	.24	53
10	.29	53
11	.30	12
12	.58	41
13	.45	12
14	.39	35
15	.53	65
16	.29	24
17	.14	18
18	.13	59
19	.25	88
20	.40	47
21	.34	59
22	-.14	18
23	.09	47
24	.43	41
25	.19	12
26	.54	59
27	.36	53
28	.62	47
29	.30	94
30	.29	53
31	-.22	47
32	.57	35
33	.00	00
34	.32	71
35	.53	65
36	.43	88
37	.27	76
38	.32	76
39	.30	94
40	-.03	35
41	-.09	59
42	.26	47
43	.14	41

Table 1. Continued

Item number	Discrimination	Difficulty
44	-.03	41
45	.22	71
46	.04	35
47	-.14	18
48	-.42	82
49	.17	71
50	.21	82
		Average = 49.8

Table 2. Summary of item analysis results for pretest, on-site (comparison) group

Item number	Discrimination	Difficulty
1	.24	33
2	.34	59
3	.36	59
4	.00	100
5	.00	33
6	-.10	37
7	.43	7
8	.32	41
9	.02	67
10	.40	33
11	.10	15
12	-.08	11
13	-.04	22
14	.09	41
15	.21	48
16	.06	11
17	.09	22
18	.46	44
19	.24	81
20	.65	19
21	.27	59
22	-.09	19
23	.43	59
24	-.09	15
25	.00	00
26	.44	22
27	.16	74
28	.33	26
29	.21	85
30	.09	41
31	-.55	56
32	.33	30
33	-.01	30
34	-.05	81
35	.24	63
36	-.06	89
37	.23	56
38	.29	59
39	.20	52
40	-.13	37
41	-.20	63
42	.06	52
43	-.12	59

Table 2. Continued

Item number	Discrimination	Difficulty
44	.51	44
45	.35	33
46	.09	22
47	-.12	19
48	.13	59
49	.49	81
50	.10	74
		Average = 44.8

Table 3. Summary of item analysis results for posttest, off-site (treatment) group

Item number	Discrimination	Difficulty
1	.27	59
2	.23	65
3	.13	53
4	.61	88
5	.20	41
6	.38	65
7	.33	53
8	.00	100
9	.73	59
10	.35	59
11	.09	71
12	.47	35
13	.36	88
14	.48	53
15	.53	65
16	.11	53
17	.23	81
18	.33	82
19	.31	94
20	.26	47
21	.38	65
22	.04	18
23	.19	47
24	.49	71
25	.11	29
26	.47	65
27	.49	76
28	.31	94
29	.31	94
30	.61	88
31	-.09	41
32	.46	47
33	-.04	24
34	.31	94
35	-.16	94
36	.52	94
37	.61	76
38	.00	100
39	.17	82
40	-.09	29
41	.09	76
42	.23	88
43	.47	82

Table 3. Continued

Item number	Discrimination	Difficulty
44	-.12	18
45	.62	53
46	.20	71
47	.32	24
48	.18	94
49	.41	88
50	.33	94
		Average = 67.5

Table 4. Summary of item analysis results for posttest, on-site (comparison) group

Item number	Discrimination	Difficulty
1	.32	74
2	.30	63
3	.26	69
4	.13	93
5	.36	59
6	.36	59
7	.15	41
8	.17	96
9	.15	67
10	-.10	78
11	.21	85
12	.20	37
13	.19	93
14	.43	58
15	.14	81
16	.26	63
17	.43	70
18	.51	63
19	-.20	89
20	.51	22
21	.31	48
22	.28	15
23	-.03	48
24	.38	70
25	.00	00
26	.50	37
27	.21	70
28	.18	81
29	.33	96
30	.45	70
31	-.46	56
32	.13	33
33	.09	37
34	.60	85
35	.10	70
36	.35	89
37	.28	70
38	.60	85
39	-.18	67
40	.16	33
41	.23	78
42	.62	85
43	.37	81

Table 4. Continued

Item number	Discrimination	Difficulty
44	.60	38
45	.16	26
46	-.04	59
47	.37	44
48	.29	96
49	.00	100
50	.37	96
		Average = 61.2

Item discrimination index indicates how well an item is able to separate scores based on knowledge of subject material. For instance, a high discrimination index would indicate that the subjects who scored well on the test tended to answer the item correctly, and the subjects who scored poorly on the test tended to answer the item incorrectly. Discrimination index on the pretest document is of less importance to the test developer, since it is assumed that there will be a considerable amount of guessing on the part of the subjects. There is no established cut-off point for item discrimination, but there is a consensus of agreement among psychometricians that the higher the discrimination index, the better, if one's purpose is to separate students by the criteria of knowledge of subject material (Brown, 1981). A negative difficulty index means that the item did not perform its function well at all. In actuality, the poorer scoring subjects tended to answer it correctly, and the better scoring subjects tended to answer it incorrectly. There were six such items on the on-site (comparison) posttest and five such items on the off-site (treatment) posttest instrument. Those items should be pulled from the test instrument and examined to determine if the problem is in the question (possibly a poor distractor or poor wording in the stem).

The reliability estimates for the pretest are not of great concern, since they were generated from scores that were predictably erratic due to lack of content knowledge on the part of the test takers. Heavy guessing on test documents tends to produce inconsistent reliability indices. However, the KR-20 reliability estimate for the posttest documents were

calculated and were found to be .65 and .77 for the on-site and off-site groups, respectively. Test publishers are generally satisfied if the reliabilities of their published tests are around the .90 mark. Teacher-made instruments rarely approach that figure (Ebel, 1965), and indices with the values obtained on the posttest documents are quite respectable.

Subjects' Performance on the Pretest/Posttest Documents

There were two groups of subjects tested: the group receiving background information off-site prior to the large-group tutorial (group 1) and the group receiving no such information (group 2). Since each group took a pretest and a posttest, the test analysis was performed for four separate test administrations: group 1 pretest, group 1 posttest, group 2 pretest, group 2 posttest, respectively. As expected, the posttests yielded higher test scores than the pretest (note Table 5).

Table 5. Comparison of test score means from the four test administrations

Group	Test	N	Raw score mean	Percent score mean	se[a]
1	pre	17	24.88	50	3.02
2	pre	27	22.44	45	3.09
1	post	17	33.18	66	2.80
2	post	27	32.22	64	2.89

[a]se = standard error of measurement.

A t-test analysis was performed on the pretest scores to determine the equivalence of group 1 and group 2 before the large-group tutorial was presented to the two groups. The results of that analysis are presented in Table 6.

Table 6. Results of t-test comparison of pretest scores for off-site (group 1) and on-site (group 2) data

Group	N	df[a]	Mean	sd[b]	se[c]	t value	2-tail prob.
1	17	42	24.88	5.10	1.24	1.91	.063
2	27		22.44	3.39	0.69		

[a]df = degrees of freedom.

[b]sd = standard deviation.

[c]se = standard error of measurement.

Examination of Table 6 yields a probability of .063, which indicates that the difference between the average pretest scores obtained by the two groups was not statistically significant (at .05 level). This leads to the conclusion that the two groups were roughly equivalent in their knowledge of the specific content material which was measured by the test instrument at that time.

Since a priori group equivalence has been indicated, it is now appropriate to compare the pretest and the posttest scores of the two groups. It is logical to assume that gain scores achieved by the two groups can be largely accounted for by cognitive gains acquired during the

large-group format tutorial administered to all subjects. A t-test analysis was performed on the pretest and posttest data to determine if the gain scores were statistically significant. The results of the t-test analysis are illustrated in Table 7.

Table 7. Results of t-test for paired samples on pretest and posttest data

Variable	N	Mean	sd[a]	se[b]	t value	2-tail prob.
Pretest	44	23.38	4.25	.64	-14.44	0.00
Posttest		32.59	5.30	.80		

[a] sd = standard deviation.

[b] se = standard error of measurement.

Examination of Table 7 yields a probability of 0.00, which indicates that the difference in average posttest scores is statistically significant (at .05 level). This leads to the conclusion that a significant gain in knowledge levels was achieved by the attendance at the large-group format tutorial. In fact, the mean scores for the entire group rose by 9.20 points, representing a gain of 39 percent over the average pretest scores.

The second question addressed by the study concerned itself with the issue of whether a pre-tutorial information packet would be helpful in

raising knowledge levels of the tutorial participants. To examine this question, a t-test analysis was performed comparing the posttest scores of group 1 and group 2. The results of the test are illustrated in Table 8.

Table 8. Results of t-test comparison of posttest scores for group 1 and group 2 subjects

Group	N	Mean	sd[a]	se[b]	df[c]	t value	2-tail prob.
1	17	33.18	5.98	1.45	42	.57	.569
2	27	32.22	4.96	0.96			

[a]sd = standard deviation.

[b]se = standard error of measurement.

[c]df = degrees of freedom.

Examination of Table 8 yields a probability of .569, which indicates that the difference between the average posttest scores is not statistically significant (at .05 level). This leads to the conclusion that the knowledge levels of both groups were roughly equivalent after the large-group format tutorial program. If the mailing of the pre-tutorial instructional packet to group 1 had a significant impact on their knowledge levels, it would be logical to assume that it would manifest itself as significantly higher posttest scores when compared with group 2. In fact, no such difference was observed. Furthermore, a case cannot be made that group 1 had significantly higher gain scores compared with group 2 due to their having a pre-tutorial packet. Examination of the gain

scores for the two groups reveals that the average of the group 1 gain scores was 8.29, compared with a 9.78 for group 2. That group had higher average gain scores when compared with group 1. Whether the difference was statistically significant was of no interest and, therefore, was not subject to statistical testing.

It was valuable to note that the gain scores for the entire group were statistically significant. It was also of value to observe that the average gain scores for each group considered separately were statistically significant. The results of the t-tests performed on each group are indicated in Table 9.

Table 9. Results of the t-test comparison of pretest and posttest scores (gain scores) for group 1 and group 2

Group	N	df[a]	Test	Mean	sd[b]	se[c]	t value	2-tail prob.
1	17	16	pre	24.88	5.10	1.24	-6.54	0.00
			post	33.18	5.98	1.45		
2	27	26	pre	22.44	3.39	0.65	-14.75	0.00
			post	32.22	4.96	0.96		

[a]df = degrees of freedom.

[b]sd = standard deviation.

[c]se = standard error of measurement.

Examination of Table 9 yields probabilities of 0.00 and 0.00, which indicates that the gain scores of both groups were statistically significant (at level .05). This finding indicates that both the group which received the pre-tutorial information packet and the group which did not receive the information benefited from attendance at the tutorial session.

The remainder of this chapter is devoted to examination of the associations among various demographic data and pretest and posttest scores.

It is logical to first consider the pretest scores and their relationship to selected demographic information. To do this, a one-way analysis of variance (ANOVA) was performed, and the results are recorded in Table 10. An ANOVA procedure was chosen because it is most appropriate when the question is whether several population means are equal.

Table 10. Results of ANOVA variable position by variable pretest scores

Position	N	Mean
Principal	4	23.75
Assistant principal	2	19.50
Department head	4	27.25
Central office	4	23.75
Teacher	30	23.03

Source	df	Sum of squares	Mean square	F ratio	F prob.
Between groups	4	94.72	23.68	1.36	.26
Within groups	39	681.72	17.48		

Examination of Table 10 indicates that despite the fact that the mean scores ranged from 19.50 to 27.25, the f-probability of .26 leads to the conclusion that there were no significant differences among the pretest scores when considered by position (at level .05). Put another way, no predictions could be made about a subject's pretest score based upon knowledge of that person's position in the educational organization.

Similarly, one-way ANOVAs were performed to investigate the possibility that grade or position level, years of experience, or education could be used to predict relative pretest scores. The results of the tests are recorded in Tables 11-13.

Table 11. Results of ANOVA, variable pretest by variable level

Level	N	Mean
Elementary	22	23.36
Middle school	9	23.44
High school	11	23.36

Source	df	Sum of squares	Mean square	F ratio	F prob.
Between groups	2	18.79	6.26	.33	.80
Within groups	39	735.86	18.87		

Table 12. Results of ANOVA, variable pretest by variable education

Group	N	Mean
BA/BS	15	22.93
BA/BS+15	4	21.00
BA/BS+30	4	22.75
BA/BS+45	1	21.00
MA/MS	7	25.00
MA/MS+15	4	22.50
MA/MS+30	4	25.25
MA/MS+45	3	28.00
PhD/EdD	2	19.50

Source	df	Sum of squares	Mean square	F ratio	F prob.
Between groups	8	162.50	20.31	1.16	.35
Within groups	35	613.93	17.54		

Table 13. Results of ANOVA, variable pretest by variable experience

Years experience	N	Mean
1-10	16	22.50
11-15	7	22.43
16-20	9	24.78
21-25	10	24.80
26-30	2	20.5
30+	0	

Source	df	Sum of squares	Mean square	F ratio	F prob.
Between groups	4	73.06	18.27	1.01	.41
Within groups	39	703.37	18.04		

Examination of the tables indicates that in all the relationships tested, it was found that there were no significant associations. Therefore, the assumption can be made that no prediction could be made about a subject's knowledge of test and measurements (as measured by the instrument) on the basis of the pretest scores when grouped by grade or position level, years of experience or education.

In an attempt to determine which demographic group benefited most from the large-group tutorial, one-way ANOVAs were performed on posttest scores to test for correlations with the same demographic data previously tested. The results are illustrated in Tables 14-17.

Table 14. Results of ANOVA, variable posttest by variable position

Position	N	Mean
Principal	4	32.00
Assistant principal	2	31.00
Department head	4	36.75
Central office	4	28.00
Teacher	30	32.83

Source	df	Sum of squares	Mean square	F ratio	F prob.
Between groups	4	161.72	40.43	1.49	.23
Within groups	39	1060.92	27.20		

Table 15. Results of ANOVA, variable posttest by variable level

Level	N	Mean
Elementary	22	33.68
Middle school	10	31.00
High school	11	32.27

Source	df	Sum of squares	Mean square	F ratio	F prob.
Between groups	2	123.74	41.25	1.47	.24
Within groups	39	1092.95	28.02		

Table 16. Results of ANOVA, variable posttest by variable education

Group	N	Mean
BA/BS	15	31.80
BA/BS+15	4	32.25
BA/BS+30	4	33.75
BA/BS+45	1	27.00
MA/MS	7	35.14
MA/MS+15	4	33.00
MA/MS+30	4	33.00
MA/MS+45	3	34.33
PhD/EdD	2	26.50

Source	df	Sum of squares	Mean square	F ratio	F prob.
Between groups	8	176.71	22.09	.74	.66
Within groups	35	1045.92	29.88		

Table 17. Results of ANOVA, variable posttest by variable experience

Years experience	N	Mean
1-10	16	31.25
11-15	7	32.14
16-20	9	35.00
21-25	10	33.00
26-30	2	27.50
30+	0	

Source	df	Sum of squares	Mean square	F ratio	F prob.
Between groups	4	151.37	37.84	1.378	.26
Within groups	39	1071.26	27.47		

Examination of the tables indicates that in all the relationships tested, there were no significant associations when posttest scores were considered by all of the above variables. This indicates that no predictions could be made about a subject's performance on the posttest based on position, years experience, grade or position level, or education.

Hypotheses Tested

In order to ascertain whether a one-day, large-group instruction tutorial could effectively raise the knowledge levels of principals in the area of student achievement measures, the following hypotheses have been tested:

Research Hypothesis I: It was operationally hypothesized that there would be significantly higher scores recorded on a posttest taken after a large-group tutorial when compared to the scores of the pretest recorded before the tutorial.

Null Hypothesis I: There will be no significant difference between pretest and posttest scores of the subjects.

To test the possibility that the subjects might have differing knowledge levels with respect to the content material measured by the test instrument before the tutorial, the following hypothesis has been tested:

Research Hypothesis II: It was operationally hypothesized that there might be significant differences among pretest scores of subjects as considered by selected demographic data.

Null Hypothesis II: There will be no significant difference in pretest scores of subjects as considered by selected demographic data.

To test the possibility that the subjects might have differing knowledge levels with respect to the content material measured by the test instrument after the tutorial, the following hypothesis has been tested:

Research Hypothesis III: It was operationally hypothesized that there might be significant differences among posttest scores of subjects as considered by selected demographic data.

Null Hypothesis III: There will be no significant difference in posttest scores of subjects as considered by selected demographic data.

To test the possibility that subjects might gain more from a large-group instruction tutorial if they were given a pre-tutorial

informational packet to study approximately two weeks in advance, the following hypothesis has been tested:

Research Hypothesis IV: It was operationally hypothesized that there might be significantly higher posttest scores recorded by a group of subjects who received a pre-tutorial informational packet when compared to a group who received no such information.

Null Hypothesis IV: There will be no significant difference in posttest scores between a group who received a pre-tutorial informational packet when compared to a group who received no such information.

The t-test results to investigate Hypothesis I are recorded in Table 7. Examination of that table yields a probability of 0.00, which indicates that the difference in the test scores is statistically significant. Since the posttests yielded higher average scores than the pretests (66 and 64 compared to 50 and 45, respectively), it is proper to maintain that the posttest scores were significantly higher than the pretest scores. The Null Hypothesis I of no significant difference is rejected, and Research Hypothesis I is retained as tenable.

The ANOVA test results to investigate Hypothesis II are recorded in Tables 10-13. In each case, it was found that there were no significant differences in pretest scores recorded by the various demographic groups. The null hypothesis is retained in each case.

The ANOVA test results to investigate Hypothesis III are recorded in Tables 14-17. In each case, it was found that there were no significant differences in posttest scores recorded by the various demographic groups. The null hypothesis is retained in each case.

The t-test results to investigate Hypothesis IV are recorded in Table 8. In this case, a probability of .569 indicates that there is no significant difference between the posttest scores of the two groups. The null hypothesis is, therefore, retained. It would seem that providing the group with background information before the tutorial did not significantly affect the subjects' cognitive gains as measured by the test instrument.

CHAPTER V. SUMMARY, CONCLUSIONS, AND RECOMMENDATIONS

Summary

The problem

The purpose of this investigation was to develop and field-test a program which could be administered to principals and staff members to raise their knowledge levels on the topic of classroom achievement testing. Operationally, the following four research hypotheses were posed:

1. Can a one-day, large-group instruction tutorial raise the knowledge levels of principals on the topic of classroom achievement testing?
2. Are there any significant differences in pretest scores when subjects are grouped according to:
   a. grade or position level,
   b. education,
   c. years experience, or
   d. position in the organization?
3. Are there any significant differences in posttest scores when subjects are grouped according to:
   a. grade or position level,
   b. education,
   c. years experience, or
   d. position in the organization?

4. Are there any significant differences in posttest scores between a group which received content information beforehand when compared to a group who received no such information?

Results - Research Hypothesis I

The question raised was: Can a one-day, large-group instruction tutorial raise the knowledge levels of principals on the topic of classroom achievement testing?

T-test comparisons of results gathered indicated that there was a statistically significant difference between the subjects' pretest and posttest scores. The null hypothesis of no differences in scores was rejected. Based on this finding, it is logical to assert that there were measurable cognitive gains on the part of the participants in the tutorial. In other words, the large-group instructional tutorial was a viable means of raising the knowledge levels of the subjects on the topic of classroom achievement testing.

Results - Research Hypothesis II

The question raised was: Are there any significant differences in pretest scores when subjects are grouped according to grade or position level, education, years of experience, or position in the organization?

A one-way analysis of variance procedure run on the results gathered revealed no statistically significant differences among scores of the subjects when they were grouped according to the question above. Based on this finding, it is logical to assert that the different demographic

groups shared roughly the same level of knowledge of classroom achievement testing (as measured by the instrument) when they came to the tutorial.

Results - Research Hypothesis III

The question raised was: Are there any significant differences in posttest scores when subjects are grouped according to grade or position level, education, years of experience, or position in the organization?

A one-way analysis of variance procedure run on results gathered revealed no statistically significant difference among posttest scores when they were grouped according to the question above. Therefore, the null hypothesis of no difference was retained. Based on this finding, it is logical to assert that the different demographic groups benefited to a similar degree by their attendance at the tutorial. Simply put, principals, teachers, supervisors, and central office staff seemed to benefit to a similar degree from their attendance at the large-group instructional tutorial on the topic of classroom achievement testing. Neither the content information, nor the mode of instruction seemed to favor one demographic group over another.

Results - Research Hypothesis IV

The question raised was: Are there any significant differences in posttest scores between a group which received content information beforehand when compared to a group which received no such information?

T-test comparisons run on the scores recorded by the group which received content information prior to the tutorial and by the group which received no content information prior to the tutorial revealed no

statistically significant difference. The null hypothesis of no difference was, therefore, retained. Based on this finding, it is logical to assert that giving background material of the nature provided to a group prior to their attendance at a large-group instruction tutorial has no measurable effect on their cognitive gains (as measured by the posttest instrument).

Conclusions

Considering the data collected and the analysis made on this investigation, the following conclusions seem warranted.

1. A one-day, large-group instruction tutorial appears to be an effective means of raising the knowledge levels of participants on the topic of classroom achievement testing.

2. Entry-level knowledge on the topic of classroom achievement testing seems to be equally distributed throughout the group of educators targeted.

3. If the audience is comprised of professional educators, the knowledge levels of all the participants regardless of years of service, education, position title, or grade or position level seem to be raised by a similar degree. Simply put, all levels of the profession seem to benefit from the experience equally.

4. Providing the participants with background material prior to a large-group tutorial of this nature seems to be of no value.

Limitations

Due to the design of this study, certain limitations must be noted at this time.

1. The limited number of participants in the study caused some cells to be small in size and, therefore, allow for only tentative conclusions with reference to certain demographic considerations.

2. The school organization chosen for the site of this study has its own test and measurement director and some of the participants have received prior inservice training on the content material in the tutorial. This training is assumed to be randomly distributed throughout the target group, and there is no way of telling what effect this prior training might have had on the test scores of the participants of the study.

3. In the limited amount of time available for the tutorial presentation, it was not possible to give the participants enough hands-on experience to get them fully acquainted with the new material and techniques.

Discussion

Although the large-group instructional tutorial was directed toward principals, a considerable number of subjects who were not principals were included in the treatment and comparison groups. This was deemed appropriate, since it was hoped that each principal would share the content material with his/her respective staff. Therefore, it was profitable to investigate how well the other staff members respond to the tutorial, since they are the eventual target of such a presentation. Furthermore, the statistical tests run to investigate *a priori* equivalence

of all demographic subgroups revealed that there were no significant differences among the performance of any of the subgroups. It is, therefore, logical to assume that the behavior of a group made up of principals alone would behave in a similar fashion.

The logistics of pretesting one group off-site and pretesting another group on-site were somewhat cumbersome. It meant that the off-site pretest group had to be taken out of the room and remain sequestered for about 45 minutes while the on-site group took their pretest. Professor Richard Manatt led the off-site group to another lecture area and presented them with information regarding research projects dealing with the topic of school improvement. After the on-site group finished their pretest, Manatt escorted the off-site group back into the lecture area and after a brief intermission, the program continued with the complete audience in attendance.

Toward the end of the presentation, a demonstration of a computerized test-making program was performed. The proprietary computer program was developed over the course of approximately five months and was designed specifically for the use of noncomputer-oriented classroom teachers. Unfortunately, the computer which was used at the tutorial site was damaged in transit and only approximately 80 percent of the program's total capabilities were demonstrated at that time. The portions of the computer program which could not be accessed at that time were explained, rather than demonstrated, by the presenter.

One of the goals of the study was to determine if giving background information to a group prior to a large-group instruction tutorial was

beneficial. The results of the study indicated that it was not. One possibility is that the members of the off-site group did not read the information given to them and, therefore, undermined its usefulness. There was no way of telling if the off-site group actually did study the material handed to them.

The analysis of the pretest and the posttest scores indicates that the subjects' level of knowledge of teacher-made tests was raised significantly by their participation in the large-group instruction tutorial. The tutorial was designed for that express purpose, so it seems to have performed its function as intended. One concern which might be raised is that the tutorial was an intensive, single episode which produced measurable results; however, do those results persist over time? That is, do the subjects retain that higher level of knowledge for a reasonable period of time after they leave the tutorial site? It is common educational wisdom that massed practice produces high gains for a short period of time, while practice distributed over time produces more long-lasting effects. Due to time and other resource constraints, it was necessary to present the tutorial as a one-day session. The tutorial package was designed as a series of four free-standing modules which could be easily presented at different sessions. These sessions could be distributed logically over a period of time with the inclusion of brief review activities. Such a method of presentation might not only help the subjects' retention of material, but it would help to reduce fatigue in the subjects which develops after extended periods of intensive training.

Another benefit of several shorter sessions is the scheduling flexibility it allows.

During a question-and-answer period after the tutorial, the researcher discussed points raised by the subjects on an informal basis. Some subjects mentioned that the material covered was totally new to them, while others indicated that they were familiar with at least parts of the material presented. Some subjects already use certain techniques and skills presented, while others indicated that they might try some of the new skills to see how well it fit into their teaching situations. It was clear that teachers in the various grade levels or subject areas were interested in different parts of the tutorial. The computerized test item bank found greater favor among teachers who tended to test often and utilized tests which had relatively large numbers of objective items. Teachers of literature, reading, writing, and others who tended to test less frequently and used essay or other forms of subjective items indicated that item banking was less useful for them.

One segment of the tutorial explored the benefits of performing a brief item analysis of test questions and student responses. This allows the instructor to locate questions which were constructed poorly and possibly misinterpreted by the students. Another benefit is that such an analysis clearly identifies subject areas where the students are especially weak, allowing the instructor to adjust the instruction accordingly. As a result of an item analysis performed on the pretest and posttest scores recorded by the participants in the study, three items were found to be deficient in their construction. Items 12, 22, and 25

(note Appendix A) were found to contain poor distractors and have been upgraded accordingly (note Appendix H).

Items 20, 44, and 47 produced unacceptably low difficulty indices, which means that they were answered incorrectly by an unusually large number of subjects. Upon close examination, the items were found to be well-constructed. This implies that the tutorial failed to teach the concepts addressed in those questions adequately. The content of the tutorial has been adjusted so that it now clarifies those concepts and emphasizes them more strongly. The next time the tutorial is presented and the subjects are tested, an item analysis will be performed to determine the value of the changes made. The cycle of presenting, testing, analyzing, and adjusting is as fundamental to the success of this program as it is to any other instructional program.

Recommendations for Practice

Although it was hoped that the tutorial would provide the scope and depth of material necessary to produce measurable cognitive gains on the part of an audience of principals, that was not its final purpose. The true function of the tutorial was to provide principals, supervisors, or mentor teachers with the skills and tools necessary to properly coach teachers to the task of producing a systematically developed, valid, and reliable classroom achievement testing program. To those ends, the tutorial was presented in such a way that it could be used whole, or in part, by another resource person. All presented materials were made available to the audience in the hopes that they will be used in accordance with their intended purpose. The materials included in the

appendices are recommended for use by both elementary and secondary principals.

Recommendations for Further Research

It is recommended that this study be replicated utilizing the tested and fine-tuned instruments developed herein. In such a future study, however, it is suggested that it take the form of an experimental design and incorporate a third (control) group. This group should take the pretest, receive no treatment, and then take the posttest. Such a study would provide a stronger link between the content of the large-group instruction tutorial and measured gain scores of the participants.

A question worth pursuing is how well universities prepare new teachers to plan, produce, and use classroom achievement measures. In order to investigate this issue, it would be necessary for a researcher to select a random sample of universities offering programs leading to educational degrees. Students from these universities could be contacted through their professors or department heads. The pre/posttest document should then be administered to these students to determine their level of knowledge on the topic of classroom achievement measures. It would not be the aim to compare universities, but rather to determine how well the contemporary student in the field of education is prepared to employ the skill of student achievement testing.

At least six recent national reports investigating the present state and possible future of educational systems in the United States note the importance of a rigorous and systematized methodology for monitoring student achievement (Almanac of National Reports, 1983). The results of

this present study indicate that sound techniques for monitoring student achievement can be taught in a meaningful and relatively efficient format to principals and teachers alike. Using the methods provided herein, principals can provide their respective staffs with the skills necessary to produce and implement a meaningful classroom testing program. Once these skills have been mastered, it may be necessary for the principal to coach and counsel the teacher in the effective use of those skills so they may become part of the teacher's normal repertoire of professional instructional tools.

BIBLIOGRAPHY

Ahmann, J. S., and Glock, M. D. Evaluating Student Progress: Principals of Tests and Measurements. 6th ed. Boston: Allyn and Bacon, 1981.

Ahrens, W., and Lehmann, I. Standardized Tests in Education. New York: Holt Rinehart and Winston, 1980.

Aiken, L. "Writing Multiple Choice Items to Measure Higher Order Objectives." Educational and Psychological Measurements 42 (Fall 1982): 803-6.

Almanac of National Reports. Publication of the National Association of Secondary School Principals. Reston, Virginia: August, 1983.

Anastasi, A. Psychological Testing. 4th ed. New York: Macmillan, 1976.

Anderson, S. B. "Take This Crash Course on Test Design." American School Board Journal (July 1981): pp. 28-30.

Ashworth, A. Testing for Continuous Achievement. London: Evans Brothers, 1982.

Austin, G. R. Process Evaluation: A Comprehensive Study of Outliers. Baltimore, Maryland: State Department of Education, 1978.

Baglin, R. "Does Nationally Norm-Referenced Really Mean Nationally." Journal of Educational Measurement 18 (Summer 1981).

Baker, F. "An Interactive Approach to Test Construction." Educational Technology 13 (March 1973): 13-15.

Berliner, D. "On Improving Teacher Effectiveness." Educational Leadership 40 (October 1982): 12-15.

Bierley, M., and Berliner, D. "The Elementary School Teacher as a Learner." Journal of Teacher Education 33 (November-December 1982): 37-40.

Bloom, B. Human Characteristics and School Learning. New York: McGraw-Hill, 1976.

Borg, W., and Gall, M. Educational Research: An Introduction. New York: Longman, 1983.

Bowers, N. "Public Reactions and Psychological Testing in Schools." Journal of School Psychology 9 (1971).

Brookover, W. School Social Systems and Student Achievement: Schools Can Make a Difference. Brooklyn, New York: Praeger Publishers, 1979.

"Brooks Eyes New Season." Eugene Register-Guard, 14 August 1981, Sec. 2, p. 1.

Brown, F. Measuring Classroom Achievement. New York: Harcourt, Brace and World, 1981.

Burns, E. The Development, Use and Abuse of Educational Tests. Springfield, Illinois: Charles C. Thomas Co., 1979.

Captrends. Publication of the Center for Performance Assessment, Northwest Regional Educational Laboratory. Portland, Oregon, 1984.

Carroll, J. "A Model of School Learning." Teachers College Record 64 (1963).

Cawelti, G. "Behavior Patterns of Effective Principals." Educational Leadership (February 1984).

Chandler, H. "In Praise of the Blackboard." Pointer 23 (Fall 1978).

Clift, J., and Imrie, B. Assessing Students, Appraising Teaching. London: Croom Helm, 1981.

Coleman, J. Equality of Educational Opportunity. Washington, D.C.: United States Office of Education, 1966.

Cooley, W. "The Instructional Dimensions Study." Educational Evaluation and Policy Analysis 2 (1980).

Cronbach, L. Essentials of Psychological Testing. New York: Harper and Row, 1970.

Cronbach, L. "Five Decades of Controversy Over Mental Testing." American Psychologist 30 (1975).

Dunstan, M. Interpretation of Item Analysis Data. Bulletin Number 5, Tertiary Education Research Center, University of New South Wales, 1969.

Durost, W., and Prescott, G. Essentials of Measurement for Teachers. New York: Harcourt, Brace and World, 1962.

Ebel, R. Measuring Educational Achievement. Englewood Cliffs, New Jersey: Prentice-Hall, 1965.

Ebel, R. "Evaluation and Educational Objectives." The Journal of Educational Measurements 10 (1973): 273-9.

Edmonds, R., and Fredrichsen, J. The Identification and Analysis of City Schools that are Instructionally Effective for Poor Children. Cambridge, Massachusetts: Harvard University Press, 1979.

Ellis, H. The Transfer of Learning. New York: Macmillan and Company, 1965.

English, F. "Curriculum Mapping." Educational Leadership 37 (April 1980): 558-9.

EPIC Diversified Systems. Teacher Self-Approved Observation System. Tucson, Arizona: Educational Innovations Press, 1970.

Fornies, F. Coaching for Improved Work Performance. New York: Van Nostrand Reinhold Company, 1978.

Friedman, M.; Brinlee, P.; and Hayes, P. Improving Teacher Education. New York: Longman Press, 1980.

Gearheart, W., and Willenberg, E. Application of Pupil Assessment Information. Denver, Colorado: Lane Publishing Company, 1974.

Glaser, R. "The Instructional Technology and the Measurement of Learning Outcomes: Some Questions." The American Psychologist 18 (August 1963).

Glaser, R., and Nitko, A. "Measurement in Learning and Instruction." In Educational Measurement. Edited by Thorndyke, R. Washington, D.C.: American Council of Education, 1971.

Goodlad, J. "A Study of Schooling: Some Findings and Hypotheses." Phi Delta Kappan 64 (March 1983): 465-70.

Gorow, F. Better Classroom Testing. San Francisco: Chandler Publishing, 1966.

Goslin, D. Teachers and Testing. Hartford, Connecticut: Connecticut Printers Incorporated, 1967.

Graeber, A. Capacity Building for a School Improvement Program, Achievement Directed Leadership. Philadelphia: Research for Better Schools, Inc. (1984).

Green, J. Introduction to Measurement and Evaluation. New York: Dodd, Mead and Company, 1970.

Green, J. Teacher-Made Tests. New York: Harper and Row, 1963.

Gronlund, N. Preparing Criterion Reference Tests for Classroom Instruction. New York: Macmillan Press, 1972.

Gross, N; Giacquinta, J.; and Bernstein, M. Implementing Organizational Innovations: A Sociological Analysis of Planned Educational Change. New York: Basic Books, 1971.

Harris, M., and Stewart, D. "Application of Classical Strategies to Criterion Reference Tests Construction: An Example." Paper presented at the annual meeting of the American Educational Research Association, New York, 1971.

Heywood, J. Assessment in Higher Education. London: J. Wiley, 1977.

Hinkle, D.; Wiersma, W.; and Jurs, S. Applied Statistics for the Behavioral Sciences. Boston: Houghton Mifflin Company, 1979.

Hively, W. Domain-Referenced Testing. Englewood Cliffs: Educational Technology Publications, 1974.

Houston, R., and Freeberg, J. "Perpetual Motion, Blindman's Bluff and Inservice Education." The Journal of Teacher Education 30 (January-February 1979): 42-49.

Hudgins, B., and Phye, G. Educational Psychology. Itasca, Illinois: F. E. Peacock Publishing Incorporated, 1983.

Hunter, M. Teach More--Faster. El Segundo, California: TIP Publications, 1976.

Jencks, C. Inequality: A Reassessment of the Effect of Family and Schooling in America. New York: Basic Books, 1972.

Joyce, B., and Showers, B. "Transfer of Training: The Contribution of Coaching." Journal of Education 163 (Spring 1981): 163-172.

Joyce, B., and Showers, B. "The Coaching of Teaching." Educational Leadership 40 (October 1982): 12-15.

Joyce, B., and Showers, B. "Improving Inservice Training: The Messages of Research." Educational Leadership 37 (February 1986): 379-385.

Kelwin, The Effective School Report, 7 January, 1984.

Klausmeier, H., and Davis, J. "Transfer of Learning." Encyclopedia of Educational Research. 4th ed. Edited by Ebel, R. Toronto, Canada: The Macmillan Company, 1969.

Le Mahiew, P. A Study of the Effects of a Program of Student Achievement Monitoring Through Testing. Ph.D. Dissertation, University of Pittsburgh, 1983.

Lezotte, L.; Edmonds, R.; and Ratner, J. A Final Report: Remedy for School Failure to Equitably Deliver Basic School Skills. East Lansing, Michigan: The Department of Urban and Metropolitan Studies, Michigan State University, 1974.

Lidz, C. Improving Assessment of Schoolchildren. San Francisco: Jossey-Bass, 1981.

Lieberman, A., and Miller, J. Staff Development: New Demands, New Realities, New Perspectives. New York: Teachers' College Press, 1979.

Lindquist, E. Educational Measurement. Washington, D.C.: American Council on Education, 1951.

Lindvall, C. Measuring Pupil Achievement and Aptitude. New York: Harcourt, Brace and World, 1967.

Marshall, J., and Hales, L. Classroom Test Construction. Reading, Massachusetts: Addison-Wesley Publishing Company, 1971.

Martuza, V. Applying Normative Reference and Criterion Referenced Measurement in Education. Boston, Massachusetts: Allyn and Bacon, 1977.

Mason, E., and Bramble, W. Understanding and Conducting Research/ Applications in Education and Behavioral Science. New York: McGraw-Hill, 1978.

Mershon, D. "An Inexpensive System for Reproducing Examinations with Minimal Typing and Proofreading." Teaching of Psychology 9 (April 1982): 108-109.

McClelland, D. "Testing for Competence Rather than Intelligence." American Psychiatrist (January 1973).

McKenna, B. "What's Wrong with Standardized Testing." Today's Education (March-April 1977): 35-38.

Morant, R. Inservice Education within the School. London, England: George, Allen and Unwin Limited, 1981.

National Commission on Excellence in Education. A Nation at Risk: The Imperative for Educational Reform. Washington, D.C.: United States Department of Education, 1983.

Nitko, A. Educational Tests and Measurements: An Introduction. New York: Harcourt Brace Jovanovich, Inc., 1983.

NSPRA. Good Schools, What Makes Them Work. Arlington, Virginia: Education USA Special Report, 1981.

O'Donnell, D. "Assessment Within Schools: A Study in One County." Educational Research 24 (November 1981): 43-48.

Popham, W. Modern Educational Measurement. Englewood Cliffs, New Jersey: Prentice-Hall Incorporated, 1981.

Popham, W. Criterion Referenced Measurement. Englewood Cliffs, New Jersey: Prentice-Hall Incorporated, 1978.

Purkey, S., and Smith, M. "Too Soon to Cheer? Synthesis of Research on Effective Schools." Educational Leadership (December 1982).

Perrone, V. The Abuses of Standardized Testing. Bloomington, Indiana: Phi Delta Kappa Educational Foundation Fastback Number 92, 1977.

"Quality Control in Curriculum Development." AASA Journal 26 (Fall 1978): 59-63.

Rebore, R. Personnel Administration in Education--A Management Approach. Englewood Cliffs, New Jersey: Prentice-Hall Incorporated, 1982.

Rogers, C. On Becoming a Person. Boston: Houghton Mifflin, 1961.

Roid, G., and Haladyna, T. A Technology for Test-Item Writing. New York: Academic Press, 1982.

Roscoe, B., and Peterson, K. "Teacher and Situational Characteristics which Enhance Learning Involvement." College Student Journal 16 (Winter 1982): 389-394.

Rosenholtz, S., and Kyle, S. "Teacher Isolation: Barrier to Professionalism." The American Educator 8 (Winter 1984): 10-15.

Rosenshine, B. "How Time is Spent in Elementary Classrooms." In Time to Learn. Edited by Denham, C., and Lieberman, A. Washington, D.C.: National Institute of Education, 1980.

Rutter, M., and Mortimer, P. Fifteen Thousand Hours: Secondary Schools and their Effects on Children. Cambridge, Massachusetts: Harvard University Press, 1979.

Schwanke, D. "Creating Conditions for Professional Practice." Journal of Teacher Education 33 (March-April 1982): 60-63.

Scriven, M. "The Methodology of Evaluation." In Curriculum Evaluation. Edited by Stake, R. Chicago: Rand McNally, 1967.

Shaw, D. "Evaluation--The Classroom Dilemma." Health Education 8 (March/April 1977): 5-6.

Shoemaker, D. "Toward a Framework for Achievement Testing." Review of Educational Research 45 (Winter 1975): 127-147.

Showers, B. Transfer of Training: The Contribution of Coaching. Oregon: Center for Education Policy and Management. University of Oregon Press, 1982.

Skinner, B. The Technology of Teaching. New York: Appleton-Century-Crofts, 1968.

Smith, F., and Adams, S. Educational Measurement for the Classroom Teacher. New York: Harper and Row, 1972.

Stanley, J., and Hopkins, K. Educational and Psychological Measurement and Evaluation. 6th ed. Englewood Cliffs, New Jersey: Prentice-Hall, 1981.

Steig, L., and Fredrick, E. School Personnel and Inservice Training Practices. West Nyack, New York: Parker Publishing Company, 1969.

Stensrood, R., and Stensrood, K. "Teaching Styles and Learning Styles of Public School Teachers." Perceptual and Motor Skills 56 (April 1983): 38-44.

Storey, A. The Measurement of Classroom Learning. Chicago: Science Research Associates (1970).

Sweeney, J. "Research Synthesis on Effective School Leadership." Educational Leadership (February 1981).

Termin, L. The Measurement of Intelligence. Boston, Massachusetts: Houghton Mifflin, 1916.

The Encyclopedia Americana, 1960 ed. S.V. "Psychology, Applied," by John F. Dashiell.

Thorndike, R., and Hagen, E. Measurement and Evaluation in Psychology and Education. New York: J. Wiley, 1977.

Tuckman, B. Measuring Educational Outcomes: Fundamentals of Testing. New York: Harcourt, Brace and World, 1975.

Willett, J. "A Meta-Analysis of Instructional Systems Applied in Science Teaching." Journal of Research in Science Teaching 20 (May 1983): 405-417.

Zavarella, J. "How to Develop a Testing Program that Reflects--Not Dictates--Your Curriculum." The National Elementary Principal (March 1980): 58-60.

Zeleny, R., and Johnson, S. The World Book of Test Taking. Chicago: The World Book Encyclopedia, 1982.

ACKNOWLEDGMENTS

Of all the pages contained in this work, this is the most difficult to compose. Not because it is difficult to offer thanks to those who have helped me along the way, rather, it is impossible to imagine any one of my family or dear friends who has not sacrificed in some way on my behalf during the course of this authorship. To those individuals, I offer my sincerest gratitude. My appreciation extends to Dr. Richard P. Manatt for his enduring patience and well-thought guidance.

The special contributions of the following persons must be noted. Thank you to Bill Matros for his computer programming talents, Dr. Kay Hoffman for her efforts in coordinating the tutorial program, Dave Peterson for his assistance with the statistical analysis, Dr. Marilyn Semones for her heroic efforts in negotiating the bureaucratic maze, and Bonnie Trede for her razor sharp secretarial skills.

Mr. Anthony Gentile and Mrs. Kathryn Gentile (mom and dad), you have helped me more than you will ever know and more than I could ever express. Sharon and Meghan, you have both served as an inexhaustible source of joy for me in this long and sometimes joyless task. Finally, to my dear wife Barbara, your patience, advice, love, and understanding are unparalleled in this universe. All this is better because of you.

This effort is dedicated in loving memory to Don-Don.

APPENDIX A.

PRETEST/POSTTEST DOCUMENT

(1-30) DIRECTIONS - These questions are of the multiple choice variety. Indicate your answers for each question by darkening in the circle on the answer sheet that corresponds with the one best response.

1. When the term "standardized" is used in reference to tests, it refers to A) how the test is administered. B) how the test is scored. C) the difficulty level of the test items. D) the arrangement of items on the page E) the item source.

2. Of the following, which is considered the most appropriate use for the results of a criterion referenced achievement test? A) selecting for fixed-quota requirements B) evaluating a specified program or curriculum C) comparing students to each other in a single classroom D) comparing students to each other between classrooms E) deciding the degree to which the test taker will benefit from instruction.

3. Which of the following terms refers to how well a test samples a domain of information? A) reliability index B) norm reference C) discrimination index D) correlation index E) content validity.

4. In the United States, the most common type of grade reporting system employed for the junior and senior high school levels utilizes which type of strategy? A) letter grades B) number or percentage C) satisfactory/unsatisfactory D) checklist rating scale E) parent conferences.

5. The results from which type of tests are most commonly perceived by principals as being most important for decisions regarding student promotion? A) standardized, norm-referenced test batteries B) published minimum competency tests C) district objective-based or continuum tests D) teacher-made curriculum tests E) aptitude scales

6. Which type of test question has the capability to sample the largest portion of a given domain? A) restricted essay B) extended essay C) short answer D) multiple choice E) subjective

7. Which type of test items are most commonly employed by teachers in the United States? A) multiple choice B) true or false C) essay D) matching E) short answer

8. Which document may be considered to be the blueprint by which a test is constructed? A) table of specifications B) curriculum mapping chart C) item analysis record D) item bank E) data record chart.

9. Of the following, the most important parameter a teacher

should consider when preparing a classroom test is A) test reliability B) mode C) the true score D) item validity E) item origin.

10. On a standard normal curve, the closest approximation of the percentage of student scores which fall within plus and minus one standard deviation about the mean is A) 95 B) 68 C) 51 D) 33 E) 16.

11. Of the following, which is considered to be the fourth level of objectives? A) global goals B) testing C) concrete behavioral objectives D) reporting to parents E) grading.

12. Which of the following is NOT included in the generally accepted definition of testing as described by Lee J. Cronbach? A) Testing should immediately follow instruction. B) Testing is a systematic procedure. C) Testing allows one to measure characteristics of people. D) Testing allows one to classify human behaviors. E) The results of tests are used to evaluate human behaviors.

13. Approximately what percentage of the average classroom curriculum is actually tested by the instructor? A) 2% B) 8% C) 40% D) 63% E) 87%

14. To classify a test as being an aptitude, attitude or achievement test, is to classify it by the A) scoring methodology B) response mode or format C) level of difficulty D) construct measured E) subject matter content.

15. In order to use the results of student achievement tests to make school personnel decisions it is necessary to A) have the instructor assist in producing the testing document. B) have a specific period set aside for testing throughout the school. C) notify the students of the intent of the test in advance of its administration. D) have the administrator assist in the production of the test. E) insure alignment between the school and test publisher's domain

16. Skinner's concept of a test as a contrived reinforcer refers to the test's ability to A) determine a student's level of mastery of subject material. B) evaluate. C) motivate. D) provide feedback to the teacher. E) reward.

17. Which type of test item is most appropriate for determining students' level of mastery of specific subject material? A) free response essay B) short answer C) objective D) extended response essay E) performance.

18. A table of specifications should be produced A) before planning B) after planning but before instruction C) after instruction but before testing D) after testing but before

grading E) after grading.

19. Which type of test attempts to measure learning which was obtained under relatively known and controlled circumstances? A) attitude surveys B) achievement measures C) aptitude scales D) interest inventories E) intelligence tests.

20. The term "criterion referenced" refers to the test's A) scoring methodology B) response mode or format C) level of difficulty D) construct measured E) origin.

21. The smallest independent unit of a testing instrument is called a (an) A) stem B) foil C) item D) artifact E) section.

22. The process of assigning numbers to an individual's performance for purposes of distinguishing among different individuals is called A) norm referencing B) evaluating C) measuring D) indexing E) recording.

23. A table of specifications is prepared to relate which three variables? A) length of test, importance of test, type of test to be administered B) content area, relative importance, length of test C) content area, target behavior, relative importance D) target behavior, relative importance of behavior, type of test to be administered E) relative importance, type of test, timing of test.

NOTE - Questions 24 and 25 have only <u>three</u> choices.

24. Which type of test is best suited for the purpose of determining the admission of a student into a quota-free program? A) achievement B) aptitude C) attitude

25. Which test classification refers to the level of difficulty of the test items? A) speed B) standardized C) cognitive.

26. Which of the following terms refers to a strategy which scores a performance against a predetermined standard of acceptable behaviors? A) norm referenced B) criterion referenced C) domain referenced D) standardized E) aligned

27. Which refers to a judgment made regarding a student's progress or level of mastery of subject material? A) measurement B) test C) reliability D) evaluation E) record

28. Which refers to a reasonable, but incorrect multiple choice response? A) stem B) foil C) deviater D) guess E) marker

29. This is a file of questions for potential use in a classroom measure. A) table of specifications B) item analysis C) item form D) item bank E) index file.

30. Which is an acceptable generalization for writing multiple choice questions? A) Use grammatical cues. B) Limit the use of "all of the above" as a response. C) Locate responses in the middle of a question. D) Utilize overlapping alternatives E) Locate repeated words or phrases at the beginning of the question.

(31-40) DIRECTIONS - These questions are of the true or false variety. On your answer sheet, indicate a <u>true</u> response by filling in circle A. Indicate a <u>false</u> response by filling in circle B. (TRUE = A, FALSE = B)

31. In order for a test to be valid, it must be reliable.

32. In order for a test to be reliable, it must be valid.

33. Compared with essay or short answer questions, objective questions are better for determining a student's depth of understanding of specific subject material.

34. In short answer or in fill-in questions, it is more desirable to locate the blanks at the end of the question that at the beginning.

35. Testing is the second level of planning objectives in a teacher's handbook.

36. Compared to letter grades, numerical grading systems are more widely employed in U.S. schools when reporting student progress.

37. Percentage of correct responses is an example of a raw score.

38. In fill-in questions, all blanks should be of equal length.

39. Criterion is another name for content area.

40. Fill-in questions are considered to be of the objective type.

41. It is appropriate to bank both objective and non-objective questions.

42. More short answer response tests tend to be given in the elementary schools than in the jr. high or sr. high schools.

43. More true or false tests tend to be given in the jr. high schools than on the elementary or the sr. high schools.

44. Alfred Benet believed that human intelligence is a fixed and quantifiable characteristic.

45. If a test results in a consistent distribution of scores

over several administrations it is considered to be valid.

46. A mastery test and a power test measure the same construct.

47. Item difficulty is based on the number of times a particular item is answered correctly. The higher the difficulty index, the more times the item is answered correctly.

48. When constructing a true or false item, it is best to present the textbook statements or definitions directly as they are written the book.

49. It is appropriate to grade students on the standard normal (bell-shaped) curve as long as there are between 25 and 30 students in the class.

50. An item analysis can be performed before a test is administered and scored.

APPENDIX B.

DEMOGRAPHIC QUESTIONNAIRE

Directions: Complete the following items on side 1 of the answer sheet by filling in the appropriate circle on the answer sheet. Use only a number 2 lead pencil.

Name--Print last name first, space between names, fill in circles beneath.

Sex--Fill in "M" or "F".

Identification Number--Fill in social security number, fill in circles below also.

Special Codes

K - Current Position
1 = Principal
2 = Assistant Principal
3 = Dean of Instruction
4 = Department Head
5 = Central Staff
6 = Other

L - Level of Assignment
1 = Elementary
2 = Middle
3 = High School
4 = Central Staff

M - Education--My most advanced degree is:
1 = BA/BS
2 = BA/BS plus 15 semester hours
3 = BA/BS plus 30 semester hours
4 = BA/BS plus 45 semester hours
5 = MA/MS
6 = MA/MS plus 15 semester hours
7 = MA/MS plus 30 semester hours
8 = MA/MS plus 45 semester hours
9 = Ph.D/Ed.D

N - Total years of experience in teaching/administration
1 = 1-10
2 = 11-15
3 = 16-20
4 = 21-25
5 = 26-30
6 = over 30

O - Years in current building assignment
1 = 1-4
2 = 5-8
3 = 9-15
4 = 16-25
5 = Over 25

APPENDIX C.

PRE-CONTACT COVER LETTER

Dear Fellow Educator,

Within the next few weeks you will be attending a presentation devoted to the topic of classroom testing. During this program you will be presented with a great deal of useful information, as well as some tools of the trade which you can apply directly in a supervisional or a classroom setting. The skills which will be taught will allow the production of more accurate and useful classroom tests while allowing the teacher to find more time to pursue other educational activities.

This program is not only useful to you, but it is also linked to a study dealing with the development of classroom testing skills. One concern of the study is to determine the a priori testing skills of the workshop participants. The most efficient way to accomplish this is with a pencil and paper test. A group of professionals has been randomly chosen to participate in this portion of the program. You have been selected as part of this group and will be identified only by ID number to ensure your anonymity, as well as the security of the testing document.

Your participation in this activity is crucial to the success of the study and you will be informed of its outcome and the fruits of your labor. Would you please take approximately forty minutes to complete the enclosed test and return it to your contact person in your school organization? When we have revised this set of instructional experiences as a result of your contribution, it will be used in School Improvement Model efforts nationwide!

Please do not discuss any part of this document or the enclosed test, as a breach of confidentiality would invalidate the work so many fellow educators have worked so hard to produce.

Thank you.

Douglas A. Gentile
Richard P. Manatt

APPENDIX D.

TEST COVER LETTER

ID#_____

Dear Participant:

This packet contains a test consisting of fifty questions (30 multiple choice and 20 true/false) on the topic of classroom testing. Please indicate your responses by filling in the appropriate circles on the bubble sheet provided using the pencils supplied. The first questions labeled "Special Codes" on the direction sheet are for statistical purposes only. Please leave nothing blank or it will invalidate the test document. An ID number has been assigned to each test paper so that no individual can be linked to his/her test. Please take approximately forty minutes or less to complete the test, and once again, be sure that nothing is left unanswered.

Thank you again. I will see you in a few weeks!

Sincerely,

Douglas A. Gentile
Iowa State University

APPENDIX E.

WORKSHOP PARTICIPANT HANDBOOK

CLASSROOM TESTING S.O.T.A.

"THE INDIVIDUAL WHO DOES WELL ON TESTS IN TODAY'S TEST-ORIENTED EDUCATION ARENA IS SOMEWHAT LIKE AN ATHLETIC HERO. PERFORMING WELL ON TESTS INCREASES THE INDIVIDUAL'S PERSONAL WORTH, RAISES EXPECTATIONS AMONG SCHOOL STAFF, AND PROVIDES NEW STATUS AT HOME AND AMONG FRIENDS. AND, GOOD PERFORMANCE EFFECTIVELY SETS ONE PERSON APART FROM ANOTHER, SERVING AS THE RITES OF PASSAGE FOR THE GOOD PERFORMER AND AS A SENTENCE TO FAILURE FOR THE POOR PERFORMER."

L.C. BECKUM

TABLE OF CONTENTS

MODULE ONE: CLASSROOM ACHIEVEMENT TESTS - FOUNDATIONS

Three general aspects of tests are as follows:

1) Tests measure observable characteristics of people.
2) Testing is limited to a finite sampling of a much larger universe of observable characteristics.
3) Testing results in a scoring strategy which allows the testor to make appropriate decisions about the target characteristics of the test-taker.

Tests may be classified by:

1) Subject matter
2) Intent
3) Administration
4) Item source
5) Scoring methodology
6) Response mode
7) Response type
8) Origin
9) Construct measured
10) Level of difficulty
11) Process measured.

Let's zero in on achievement tests. Achievement tests purport to measure learning obtained under relatively controlled circumstances. General uses are:

1) Determining student's level of mastery
2) Determining student's entry level skills
3) Diagnosing deficiencies in learning
4) Certifying for competency
5) Admitting to quota-free programs
6) Measuring growth or progress in a subject
7) Aligning curriculum, evaluating program
8) Contributing to personnel decisions

Teachers use classroom tests for three reasons:

1) Evaluation
2) Motivation
3) Feedback

MODULE TWO: ITEM TYPES

In testing jargon, an item is simply an individual test question. Objective test items also include possible correct responses.

Item types include:

1) Matching
2) Multiple choice
3) True/false
4) True/false with corrections
5) Short answer
6) Fill-in
7) Essay (restricted)
8) Essay (extended)
9) Performance tests - observations, interviews, projects, extended assignments, laboratory exercises

On the next page you will find generalizations for writing various item types.

On the following page you will find a series of twenty statements describing the possible advantages of the various item types. In the space provided at the right of each statement, fill in either an E (essay), SA (short answer), or O (objective), depending upon which item type you feel BEST fits the description provided.

GENERALIZATIONS FOR WRITING ITEM TYPES

MULTIPLE CHOICE QUESTIONS

1. Avoid grammatical cues.
2. Avoid overlapping alternatives.
3. Keep it simple - NO window dressing.
4. Ask direct, complete and unambiguous questions.
5. Locate the alternatives at the end of the question.
6. Underline negatives or other delimiters. (ex. most, least)
7. Locate repeated words or phrases at the end of the stem.
8. Avoid worthless or 'throw away' foils.
9. Base each alternative on a single central problem.
10. List alternatives logically, if possible.
11. Limit use of 'none of these', or 'all of these'.

ESSAY TYPE QUESTIONS

1. Use when other question types are NOT suitable.
2. List any delimitations or specifications clearly.
3. Indicate point allotment and criterion for grading.

MATCHING TYPE QUESTIONS

1. Use homogeneous premises and homogeneous responses.
2. Use a greater number or responses than premises.
3. Indicate that responses are re-usable if appropriate.
4. Use longer, more complex statements as premises.
5. Use shorter, less complex statements as responses.
6. Keep exercises relatively short (6-8 items).
7. Locate entire matching exercise on the same page.

TRUE/FALSE TYPE QUESTIONS

1. Items should be definitely true or definitely false.
2. Avoid copying textbook statements or definitions.
3. Avoid terms such as never, always and sometimes.
4. Limit use of negatives.
5. Underline delimiters (ex. former, next, least).

SHORT ANSWER, FILL-IN TYPE QUESTIONS

1. Elicit direct short answers.
2. Items should be unambiguous and elicit specific responses.
3. Locate the blanks (if any) toward or at the end of the item.
4. All blanks should be of equal length.
5. Avoid verbal and punctuation cues.

ESSAY- E
SHORT ANSWER- SA
OBJECTIVE- O

ADVANTAGES OF VARIOUS ITEM TYPES

1.	SAMPLES A LARGE PORTION OF THE DOMAIN	1. _____
2.	MEASURES ORIGINALITY, INNOVATION	2. _____
3.	MEASURES ABILITY TO EXPRESS THOUGHTS LOGICALLY	3. _____
4.	MEASURES RECALL	4. _____
5.	MEASURES ABILITY TO SYNTHESIZE	5. _____
6.	MEASURES ABILITY TO ANALYZE NEW SITUATIONS	6. _____
7.	DIAGNOSES LEARNING DEFICIENCIES	7. _____
8.	ISOLATES LEARNING FROM OTHER VARIABLES (SPELLING, HANDWRITING, WORD USAGE, NEATNESS)	8. _____
9.	ITEMS CAN BE QUICKLY PREPARED	9. _____
10.	ITEMS CAN BE QUICKLY SCORED	10. _____
11.	ITEMS CAN BE CONSISTENTLY SCORED (OVER TIME AND FROM ONE SCORER TO ANOTHER)	11. _____
12.	LIMITS HALO/PITCHFORK EFFECT	12. _____
13.	LIMITS GUESSING	13. _____
14.	LIMITS CHEATING, COPYING	14. _____
15.	ITEMS CAN BE BANKED	15. _____
16.	PARALLEL TESTS MAY BE CONVENIENTLY PRODUCED	16. _____
17.	MAY DETERMINE LEVEL OF MASTERY OF SPECIFIC MATERIAL	17. _____
18.	MEASURES ABILITY TO SUPPORT A POSITION	18. _____
19.	MEASURES ABILITY TO MAKE FINE DESCRIMINATIONS AMONG REASONABLE CHOICES OR ALTERNATIVES	19. _____
20.	MEASURES A NARROW DOMAIN IN-DEPTH	20. _____

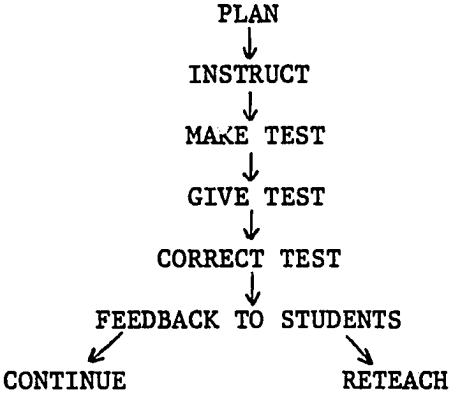
MODULE THREE: TESTING AS A PLANNING OBJECTIVE

When teachers plan, they generally proceed from global goals, to concrete objectives and finally to specific methodologies, materials and timelines.

Testing experts maintain that the test is actually the fourth and most clearly specified level of objectives since it is the behavioral target at which instruction is aimed.

On the following pages you will find a list of illustrative verbs which may prove useful in writing behavioral objectives in the cognitive domain. These same terms may also be utilized when writing test items to those same objectives.

The normal instructional process can be diagrammed in the following way:



ILLUSTRATIVE VERBS FOR USE IN WRITING OBJECTIVES AND TEST QUESTIONS IN THE COGNITIVE DOMAIN

KNOWLEDGE

Acquire	Indicate	Outline	Recite	Select
Count	Label	Point	Recognize	State
Define	List	Quote	Record	Tabulate
Distinguish	Match	Read	Repeat	Trace
Draw	Name	Recall	Reproduce	Write
Identify				

COMPREHENSION

Associate	Differentiate	Extrapolate	Illustrate	Reorder
Change	Discuss	Fill in	Interpret	Represent
Conclude	Distinguish	Generalize	Paraphrase	Restate
Compare	Draw	Give in own	Predict	Rewrite
Contrast	Estimate	words	Prepare	Summarize
Convert	Explain	Give examples	Read	Transform
Describe	Extend	Infer	Rearrange	Translate
Determine	Interpolate			

APPLICATION

Apply	Demonstrate	Illustrate	Predict	Show
Calculate	Develop	Manipulate	Prepare	Solve
Choose	Discover	Modify	Produce	Transfer
Classify	Employ	Operate	Relate	Use
Complete	Examine	Organize	Restructure	Utilize
Compute	Generalize	Practice		

ANALYSIS

Analyze	Contrast	Discriminate	Infer	Relate
Break down	Deduce	Distinguish	Order	Select
Categorize	Deduct	Group	Outline	Separate
Classify	Diagram	Identify	Point out	Subdivide
Compare	Differentiate	Illustrate	Recognize	Transform

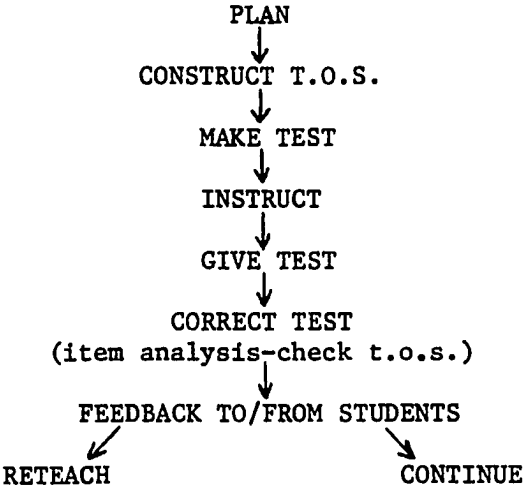
SYNTHESIS

Arrange	Deduce	Generalize	Prescribe	Rewrite
Categorize	Derive	Generate	Produce	Specify
Combine	Design	Integrate	Propose	Summarize
Compile	Devise	Modify	Rearrange	Synthesize
Compose	Develop	Originate	Reconstruct	Tell
Constitute	Document	Organize	Relate	Transmit
Construct	Explain	Plan	Reorganize	Write
Create	Formulate	Prepare	Revise	

EVALUATION

Appraise	Criticize	Distinguish	Measure	Standardize
Argue	Critique	Evaluate	Rank	Summarize
Assess	Decide	Grade	Rate	Support
Compare	Describe	Judge	Recommend	Test
Conclude	Determine	Justify	Relate	Validate
Consider	Discriminate	Interpret	Select	Verify
Contrast				

Assuming that testing is an integral part of the planning process, the flow diagram should look more like this:



A table of specifications (or t.o.s.), is a blueprint for both teaching and testing. In one simple form it consists of a table indicating the amount of importance, or possibly the time allotment, attached to specific skills in each content area.

On the next page you will find a completed t.o.s. which will act to illustrate its use. On the following page you will find a blank t.o.s. which you may use as a template or guide in producing your own tables of specifications.

It is important to understand that although the test is compiled prior to instruction, it does not follow that one teaches to the test. On the contrary, it is NOT advisable to do so. That practice severely limits the useful scope of instruction and it freezes teaching from topic to topic. The true value of a t.o.s. is both to allow the production of a test which reflects instruction and to set a pattern or blueprint for that instruction. Therefore, it IS recommended that an instructor teaches to the t.o.s. and that he/she treats it as a functional lesson plan.

TOPIC - STOICHIOMETRY - CHAPTER 5

CONTENT AREAS / SKILLS -	KNOWL.	COMP.	APP.	ANAL.	SYN.	EVAL.	C.A. TOT.
MOLES	2	4	6	2		2	16 / 28 %
DENSITY	1	1	3				5 / 9 %
EQUATIONS	1	1	4	4	3	2	15 / 27 %
PERCENT COMP.	3	2	3	5			13 / 23 %
SOLUTIONS	1	2				3	6 / 11 %
SKILL TOTAL	8	10	16	11	3	7	55 TEST TOTAL

TOPIC - _____

SKILLS -	KNOWL.	COMP.	APP.	ANAL.	SYN.	EVAL.	C.A. TOT.
CONTENT AREAS							____ %
							____ %
							____ %
							____ %
							____ %
SKILL TOTAL							TEST TOTAL

MODULE FOUR: MAKING THE TEST

Now comes the easy part! Once the planning process is completed, and a t.o.s. has been assembled, production of a valid test becomes a routine matter.

The systematic production of a classroom achievement test is most easily accomplished by the maintenance of an item bank. Item banks consist of a relatively large number of test items filed under specific content areas. Each item may have several parallel forms for use in making alternate test instruments, pre-tests and even practice exercises for use during instruction.

Test files have traditionally been kept on 3 x 5 cards in a card box or filing cabinet. They are effective, but rather cumbersome in actual use. Present computer technology allows for a much more streamlined system to store and retrieve test items. The added bonus using this system is that it types a neat, error-free test master and a test key for the instructor. Given access to the proper computer and software, questions can be filed under specific topic labels and even coded with reference to behavioral objective to be tested. Any type of question can be banked, and production of a test is a simple matter of calling up the desired items by category and selecting among them. It is also possible to easily and quickly update the item bank by adding or deleting questions as seen fit by the instructor.

GLOSSARY OF TERMS COMMONLY EMPLOYED IN PSYCHOMETRICS

ACHIEVEMENT TEST - A test designed to measure a student's level of mastery of a body of knowledge or proficiency in certain skills obtained under relatively known circumstances.

APTITUDE TEST - A test designed to measure a student's potential for development along a specified line by measuring knowledge or proficiency in certain skills obtained under relatively uncontrolled or unknown circumstances.

ATTITUDE SCALE (TEST) - A testing device designed to assess an individual's position on some object or proposition relative to three components; affective, cognitive and behavioral.

BEST ANSWER ITEM - A multiple choice item in which the foils or distracters are not totally wrong. This type of item often requires the test taker to make relatively fine distinctions among the possible choices, thereby utilizing thinking skills on a higher order than pure recall.

CONTRIVED REINFORCER - Term used by B.F. Skinner to describe teacher's use of classroom tests as short range targets for students' learning.

CONVERTED SCORE - A test score which has been changed from the raw score form as dictated by a predetermined mathematical formula or relationship. For example, percentages are converted scores, as are stanines.

CORRELATION COEFFICIENT - A number between plus and minus one which is mathematically derived and indicates the degree of relationship between two or more variables or scores.

CRITERION REFERENCE TEST (CRT) - A testing device which is meant to be scored against a predetermined standard of competent behaviors.

DIFFICULTY INDEX - A scale intended to represent the relative ease of a specific test item. The number is mathematically derived, with reference to the percentage of students responding correctly to the item.

DISCRIMINATION POWER - The ability of a test item to separate individuals based on their behaviors with reference to a specific construct or domain.

DOMAIN - The body of knowledge or behaviors on which instruction and learning is based. Sometimes loosely referred to as content area.

DOMAIN REFERENCED TEST (DRT) - A testing device whose content is derived from a specified body of knowledge or behaviors.

ESSAY TEST - A testing devise consisting of one or more items and a list of instructions which require the test taker to compose a more or less original, extended written response. This type of test is used to determine a student's ability to put forth and defend a proposition logically and persuasively. Essay tests also allow the instructor to measure grammar, spelling and handwriting.

EVALUATION - A judgement of merit, appropriately based on measurements and a synthesis of other valid evidence.

FOIL (DISTRACTER) - An incorrect response option used in multiple choice items.

FREE RESPONSE ITEM - An item, usually essay type, in which the content of the response is left solely to the discretion of the test- taker.

ITEM - The smallest independent unit of a testing instrument. Items may or may not contain a list of possible responses depending upon the objectives of the measure.

ITEM ANALYSIS - The process of testing the items in a measure with regard to such constructs as validity, reliability, difficulty, and discrimination power. Normally, for teacher-made classroom measures the greater importance is attributed to information regarding validity and item difficulty.

ITEM BANK - A systematized filing of a large number of possible test items for subsequent use in testing and instruction.

MASTERY TEST - See power test.

MEAN - The arithmetic average of a set of scores.

MEASUREMENT - The process of assigning numbers or labels to a group for the purpose of distinguishing among them on the basis of specified characteristics.

MEDIAN - The point in a score distribution which divides it into two equal parts.

MODE - The most frequently occurring score in a set of scores.

NORM REFERENCED TEST (NRT) - A testing device which is meant to be scored in such a manner that each score in the distribution can be statistically compared to every other score. Such a distribution has a predetermined mean and standard deviation. The scores are mathematically fit to a typical bell-shaped curve

in which approximately 68% of the scores are contained within plus and minus one standard deviation around the mean. Additionally, 95% of the scores and 99% of the scores are contained within plus and minus two and three standard deviations respectively.

OBJECTIVE TEST - A testing device in which each item is supplied with a predetermined set of possible responses, so that subjective opinions or judgments in scoring are minimized. This type of test may measure all levels of thinking independently of the grammatical, handwriting and spelling skills possessed by the test taker.

POWER TEST - A measuring device consisting of a moderate number of items starting with those of low difficulty and progressing to items of high difficulty. The objective of such a test is to determine the test taker's depth of understanding of the subject material. Usually a generous amount of time is allotted for completion of the test so that speed as a scoring factor is minimized.

PSYCHOMETRICS - The science of measuring (testing) and evaluating.

RANDOM SAMPLE - A sample selected in such a way as to guarantee equal probability of selection to all possible samples of the same size which could be chosen from the universe in question. In order for the sample to be truly random, the universe must be made up of a relatively large number of members. Experts maintain that the minimum number of members should be in the 125 to 150 range. For this reason as well as others dealing with the concept of group self-selection, it is generally agreed that obtaining a true random sampling within a classroom setting is not possible.

RAW SCORE - The total number of correct responses a student achieves on a measure.

RELIABILITY - The consistency of a test's results. This may be measured for a single test over time, two equivalent forms of the same test, or for two halves of a single test. It is usually expressed as a correlation coefficient.

SIGNIFICANT DIFFERENCE - A large enough difference between two comparable statistics such that the mathematically defined probability that such a difference may be attributable to pure chance is less than some predefined limit.

SPEED TEST - A measuring device which contains a relatively large number of items having a relatively low difficulty. The main objective is to determine how quickly the test taker can display certain skills with little regard to the depth of knowledge of

material.

STANDARD DEVIATION - A mathematically derived unit of dispersion of scores about the mean.

STANDARDIZED - Referring to the methodology of test administration, standardized test are designed to be administered within strict guidelines. Such parameters as time allotment, response mode, and even the type of writing utensil and answer sheets used are among those things commonly delineated.

STEM - The interrogative portion of an item.

TABLE OF SPECIFICATIONS (T.O.S.) - A tabular representation relating three variables; content area, target behavior or skill, and relative importance. When used for instruction, this importance may be indicated as a time allotment. When used for test construction this importance may indicate the total number of questions alloted to that portion of the topic.

TRUE SCORE - A derived score which is considered an error-free score for a particular person on a particular test. For example, corrections are mathematically made for guessing, mismarking answers and other miscellaneous random errors.

VALIDITY - How well a test measures what it purports to measure. Sometimes referred to as the truthfulness of a test.

APPENDIX F.

HANDBOOK COVER LETTER

CONGRATULATIONS!

You have contributed to your profession in a way in which few people can claim. Please accept this workshop handbook and read through its contents. To do so will not only prepare you for the conference, but may offer you some information which you will find immediately useful in your position. Once again, this is for your personal use, and please do not share its contents with others until after the workshop, since this will affect the study in a negative way. The other members of the group will receive this information on-site.

Thank you once again,

Douglas A. Gentile
Iowa State University

APPENDIX G.

OPTICAL SCANNING SHEET

SIDE 1

GENERAL PURPOSE - NCS® - ANSWER SHEET

FOR USE WITH ALL NCS SENTRY OPTICAL MARK READING SYSTEMS
SEE IMPORTANT MARKING INSTRUCTIONS ON SIDE 2

NAME (Last, First, M.I.)

SEX (M) (F)

GRADE OR EDUC 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16

BIRTH DATE			IDENTIFICATION NUMBER										SPECIAL CODES					
MO.	DAY	YR.	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
Jan.																		
Feb.																		
Mar.	0 0	0 0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Apr.	1 1	1 1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
May	2 2	2 2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
Jun.	3 3	3 3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
Jul.	4 4	4 4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
Aug.	5 5	5 5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5
Sep.	6 6	6 6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6
Oct.	7 7	7 7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7
Nov.	8 8	8 8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8
Dec.	9 9	9 9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9

1 A B C D E	11 A B C D E	21 A B C D E	31 A B C D E	41 A B C D E
2 A B C D E	12 A B C D E	22 A B C D E	32 A B C D E	42 A B C D E
3 A B C D E	13 A B C D E	23 A B C D E	33 A B C D E	43 A B C D E
4 A B C D E	14 A B C D E	24 A B C D E	34 A B C D E	44 A B C D E
5 A B C D E	15 A B C D E	25 A B C D E	35 A B C D E	45 A B C D E
6 A B C D E	16 A B C D E	26 A B C D E	36 A B C D E	46 A B C D E
7 A B C D E	17 A B C D E	27 A B C D E	37 A B C D E	47 A B C D E
8 A B C D E	18 A B C D E	28 A B C D E	38 A B C D E	48 A B C D E
9 A B C D E	19 A B C D E	29 A B C D E	39 A B C D E	49 A B C D E
10 A B C D E	20 A B C D E	30 A B C D E	40 A B C D E	50 A B C D E

51 A B C D E	61 A B C D E	71 A B C D E	81 A B C D E	91 A B C D E
52 A B C D E	62 A B C D E	72 A B C D E	82 A B C D E	92 A B C D E
53 A B C D E	63 A B C D E	73 A B C D E	83 A B C D E	93 A B C D E
54 A B C D E	64 A B C D E	74 A B C D E	84 A B C D E	94 A B C D E
55 A B C D E	65 A B C D E	75 A B C D E	85 A B C D E	95 A B C D E
56 A B C D E	66 A B C D E	76 A B C D E	86 A B C D E	96 A B C D E
57 A B C D E	67 A B C D E	77 A B C D E	87 A B C D E	97 A B C D E
58 A B C D E	68 A B C D E	78 A B C D E	88 A B C D E	98 A B C D E
59 A B C D E	69 A B C D E	79 A B C D E	89 A B C D E	99 A B C D E
60 A B C D E	70 A B C D E	80 A B C D E	90 A B C D E	100 A B C D E

SIDE 2

101 A B C D E	111 A B C D E	121 A B C D E	131 A B C D E	141 A B C D E
102 A B C D E	112 A B C D E	122 A B C D E	132 A B C D E	142 A B C D E
103 A B C D E	113 A B C D E	123 A B C D E	133 A B C D E	143 A B C D E
104 A B C D E	114 A B C D E	124 A B C D E	134 A B C D E	144 A B C D E
105 A B C D E	115 A B C D E	125 A B C D E	135 A B C D E	145 A B C D E
106 A B C D E	116 A B C D E	126 A B C D E	136 A B C D E	146 A B C D E
107 A B C D E	117 A B C D E	127 A B C D E	137 A B C D E	147 A B C D E
108 A B C D E	118 A B C D E	128 A B C D E	138 A B C D E	148 A B C D E
109 A B C D E	119 A B C D E	129 A B C D E	139 A B C D E	149 A B C D E
110 A B C D E	120 A B C D E	130 A B C D E	140 A B C D E	150 A B C D E

151 A B C D E	161 A B C D E	171 A B C D E	181 A B C D E	191 A B C D E
152 A B C D E	162 A B C D E	172 A B C D E	182 A B C D E	192 A B C D E
153 A B C D E	163 A B C D E	173 A B C D E	183 A B C D E	193 A B C D E
154 A B C D E	164 A B C D E	174 A B C D E	184 A B C D E	194 A B C D E
155 A B C D E	165 A B C D E	175 A B C D E	185 A B C D E	195 A B C D E
156 A B C D E	166 A B C D E	176 A B C D E	186 A B C D E	196 A B C D E
157 A B C D E	167 A B C D E	177 A B C D E	187 A B C D E	197 A B C D E
158 A B C D E	168 A B C D E	178 A B C D E	188 A B C D E	198 A B C D E
159 A B C D E	169 A B C D E	179 A B C D E	189 A B C D E	199 A B C D E
160 A B C D E	170 A B C D E	180 A B C D E	190 A B C D E	200 A B C D E

NCS Trans-Optic® 4521 100999897 A2804-4 Printed in U.S.A.

GENERAL PURPOSE

NCS®

ANSWER SHEET

FOR USE WITH ALL NCS SENTRY OPTICAL MARK READING SYSTEMS

EXAMPLES	IMPORTANT DIRECTIONS FOR MARKING ANSWERS
WRONG 1 ① ⊗ ③ ④ ⑤ WRONG 2 ① ② ⊘ ④ ⑤ WRONG 3 ① ② ③ ◐ ⑤ RIGHT 4 ① ② ③ ● ⑤	• Use #2 pencil only. • Do NOT use ink or ballpoint pens. • Make heavy black marks that fill the circle completely. • Erase cleanly any answer you wish to change • Make no stray marks on the answer sheet.

DO NOT WRITE IN THIS SPACE

APPENDIX H.

TEST ITEM CORRECTIONS

12. Which of the following is NOT included in the generally accepted definition of testing as described by Lee J. Cronbach? A) All tests should be of the objective variety. B) Testing is a systematic procedure. C) Testing allows one to measure characteristics of people. D) Testing allows one to classify human behavior. E) The results of tests are used to evaluate human behaviors.

22. The process of assigning numbers to an individual's performance for purposes of distinguishing among different individuals is called A) computing B) evaluating C) measuring D) stepping E) recording.

25. Which test classification refers to the level of difficulty of the test items? A) speed B) standardized C) attitude