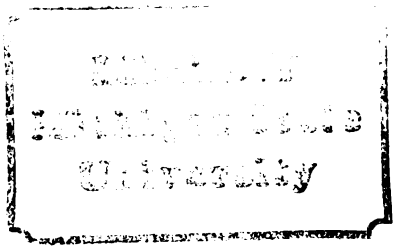


This is to certify that the

thesis entitled

A Study of the Qualifications of the Teacher as an Evaluator in Bangkok, Thailand

presented by

Tuanjai Sethtasakko

has been accepted towards fulfillment
of the requirements for

Ph.D. degree in Education:
Measurement, Evaluation & Research Design

William Mehrens

Major professor

Date October 14, 1980

A STUDY OF THE QUALIFICATIONS OF THE TEACHER AS AN EVALUATOR IN BANGKOK, THAILAND

By

Tuanjai Sethtasakko

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

Department of Counseling, and Educational Psychology

1980

ABSTRACT

A STUDY OF THE QUALIFICATIONS OF THE TEACHER AS AN EVALUATOR IN BANGKOK, THAILAND

By

Tuanjai Sethtasakko

This study was aimed at providing data concerning the quality of the teacher as an evaluator to the administrators and educators in Thailand. It was the purpose of the study to find out which groups of teachers actually need in-service training and in which areas of measurement the need is the greatest. This study also yields some follow-up information on the effects of the previous in-service programs in measurement and the effects of a measurement course offered by the teacher-training institutions.

The population of interest of this study was the public elementary and secondary school teachers, who are under the Ministry of Education, in Bangkok. The instrument used in the study was a questionnaire concerning the teachers' opinions on national testing and their perceived needs in measurement and a true-false test measuring basic knowledge in measurement corresponding to the four subject matters that the teachers should know. There were eight questionnaire items and fifty-two test items, thirteen items for each subscale. The instrument was sent to 540 Thai teachers who were randomly selected from 12 strata. The stratification was based on the three variables: level of school - elementary school or secondary school; level of teacher education - teaching

certificate holders or bachelor degree holders; and teaching experience - less than or equal to three years, between four to ten years, or more than ten years.

The design of the study was a 2 x 2 x 3 factorial design with four measures. The four areas of basic knowledge in measurement and evaluation were: planning a classroom test; item writing; item analysis; and test score statistics and marking system. The design was crossed and balanced with 30 observations per cell. The multivariate repeated measures analysis was employed to test the fifteen null hypotheses.

It was found that the subject matter by teacher education interaction was significant. The analyses showed that both certificate teachers and degree teachers got their lowest scores on Planning a Classroom Test subscale. The degree teachers got their highest scores on Test Score Statistics and Marking System subscale, but the certificate teachers got their highest scores on the Item Analysis subscale. The degree teachers got higher scores than the certificate teachers in all subscales.

An interaction effect was also found on the level of school by teacher education interaction. The total score mean indicated that the degree teachers in secondary schools got higher scores than the degree teachers in elementary schools, but the mean of certificate teachers in secondary schools was slightly lower than the mean of certificate teachers in elementary schools. There was no significant difference between certificate elementary school teachers and certificate secondary

school teachers, but a significant difference between degree elementary school teachers and degree secondary school teachers was found at the .05 level.

Further analyses were done to compare measurement needs among various groups of teachers, to observe the relationship between perceived needs and measurement needs, and to compare the mean differences between sample means and criterion scores. It was found that the teachers who took a college measurement course had less need for instruction on measurement than those who did not take a course. There were no significant differences on measurement needs among teachers who attended the training program and those who did not attend the program, or among teachers who favored and did not favor the national testing program. No relationship between perceived needs and measurement needs was found.

There were significant differences between sample means and criterion scores both on total mean and subscale means. The results showed that the teachers in Bangkok had measurement needs in all four subject matter areas.

ACKNOWLEDGEMENTS

I wish to express my sincere appreciation to those persons who have assisted so greatly in conducting this dissertation and doctoral study.

First, to the dissertation director and chairman of the doctoral committee, Dr. W. A. Mehrens, whose interest, assistance, and guidance were essential to the development and completion of the study.

Second, to the other members of the doctoral committee who made important contributions: to Drs. R. L. Ebel and S. Cherney for their help in the development and improvement of the study; to Dr. V. Scheifley for her help in analyzing the data and her moral support during the past few years.

Third, to Mr. Boonlue Tong-Yoo and Ms. Orathai Tong-Yoo for their help in distributing and collecting the questionnaire. To the P.E.O. International Peace Scholarship Fund and the Sage Foundation Fund for some financial aid in the preparation of this dissertation.

Finally, to my parents, Mr. and Mrs. Kamol Sethtasakko, and my husband, Dr. Saichol Ketsa, for their support, encouragement and patience throughout the doctoral study.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

THE PROBLEM

Introduction

From the earliest beginnings of society, people have measured the abilities of other people and have recognized the existence of differences in the abilities possessed by different individuals. Impressions of people often develop from unsystematic observations. Often these impressions are totally incorrect or unfair. The drawing of conclusions about students from classroom examination results is analogous to the drawing of everyday life impressions from social incidents. Here too, conclusions drawn are sometimes radically incorrect or unfair to the student who has taken the examination.

What causes these misinterpretations? While the interpreter is sometimes to blame, the major cause is usually the test: it does not properly measure the ability being judged; it does not adequately sample the individual's behavior; it does not offer an accurate or consistent measure of the trait; or the test is not sensitive enough to measure small gradations in ability.

Tests and testing processes are more closely related to everyday activity than is commonly realized. The situational conditions which lead to accurate impressions of human behavior are much the same as the

characteristics of a classroom test which lead to an accurate evaluation of student achievement. In fact, such conditions are essential to effective measurement throughout the entire field of evaluation.

Tests have little or no value in their own right. They are good or bad primarily in terms of how they are used to affect the learner. Tests can improve the effectiveness of the instructional decisions by providing more objective information on which to base the judgments. The use of tests can have an immediate and direct effect on the learning of students.

Tests have become an integral part of our life and many decisions are often made on the basis of a student's score. It is very essential to recognize that test results may play an important role in students' lives.

If questions are well-chosen and tests well-constructed, they can help students learn how to organize, analyze, and judge ideas and concepts; how to sort relevant details from irrelevant ones; and how to think critically about the possible relationships in the materials they have studied. Students who are given varied and challenging tests receive, in effect, valuable learning experiences. The tests are also likely to enable them to generalize about the importance of a course or even the value of education itself.

Thus, a teacher's test is a powerful contributor to what students learn and how they learn it.

Need for the Study

Teachers have an obligation to provide their students with the best instruction possible. This implies that they must have some procedures

whereby they can reliably and validly evaluate how effectively their students have been taught. The need to employ all applicable techniques of appraisal in the schools is a practical matter faced daily by teachers.

One of the major responsibilities given to the teacher is the difficult, but necessary, task of assigning grades. Considering the importance of grades to the students, one must be as certain as he can be that this is done wisely. Grades should reflect both achievement and quality of performance. A high or low grade may be a determining factor in finding a job, being permitted to sit for university entrance examinations, or establishing individual interest in certain careers and vocations. Examinations help to determine not only the degree of achievement but also the individual's achievement relative to others in his class. For all these reasons, evaluative instruments must be as reliable and valid as they can be because they will provide the basic measurements from which the final grade is made.

Teachers must become proficient in constructing classroom tests because they occupy the central role in the evaluation process. Unfortunately, very often teachers have not been given adequate preparation in this area of competence. There is evidence that the problem of preparing teachers in measurement and evaluation is real, and substantial. Many teachers do not seem to be adequately prepared in this respect (Goslin, 1967; Roeder, 1972).

The quality of teacher-made tests in Thailand is critical because of two important reasons. First if a student fails the final examination at the end of the school year he has to be in the same grade for another year with some unfavorable attitude and with the label "repeater."

Secondly, teacher-made tests are not only used for the final examinations, they are also used for the entrance examinations. Since there are limited seats for students in secondary school and the university level, entrance examinations play a very important role in selecting students. Students who want to get into secondary schools or universities must obtain high percentile ranks on the entrance examinations for admission. Thus, the need for improving the competence of teachers in measurement and evaluation is urgent.

It is recognized that the instructional values of teachers' roles in testing and evaluation in elementary and secondary school in Thailand has received much less attention. If the advantages of objective testing are to be fully realized, it seems clear that the teacher training institutions will have to pay increased attention to the problems of test construction as they apply to classroom teachers and provide the future teachers with the kinds of skills they will need to do an adequate job of constructing tests. The teacher training institutions might perform a particularly useful role in helping the teachers by developing practical courses in test construction.

There is little doubt that examination and evaluation have become an integral part of our academic life. Although crucial decisions are often made on the basis of student's test scores, little if any thought is given to the qualifications or skill of the teacher as an evaluator.

Purpose of the Study

The purpose of this study was to identify the areas of instruction in measurement which are needed most by the elementary and secondary

school teachers in Bangkok, Thailand, and also to compare the basic knowledge in measurement and evaluation of teachers who are grouped according to the amount of their experience in teaching, to the levels of their education, and to the levels of school they teach. Measurement needs in four fundamental areas in measurement and evaluation are measured: planning a classroom test, item writing, item analysis, and test score statistics and marking system.

Secondly, the study tested the relationship between measurement needs and perceived needs of teachers in Bangkok. It was also aimed at providing the data and information concerning the quality of Thai teacher as an evaluator to the administrators and educators in teacher training institutions in Thailand.

Statement of the Problem

The research described here was an investigation of the following questions:

1. Is there a difference in measurement needs between elementary school teachers and secondary school teachers in Bangkok?
2. Do teachers who have higher levels of education have less measurement needs than those who have lower levels of education?
3. Is there any difference in measurement needs between teachers who have more teaching experience and those who have less teaching experience?
4. Is there any difference in measurement needs among the four subject matters covered in the study?

The result of this study will provide information about the qualification of the teacher as an evaluator to the administrators and educators in Thailand. As a result of this study, one should be able to conclude: a) which group of teachers in Bangkok has the most urgent need for in-service programs in measurement and evaluation, and b) what area of in-service training is most needed.

Limitation of the Study

This study was based on the sample of elementary and secondary school teachers who have been teaching in public school in Bangkok. The sample includes only the teachers who are under the Ministry of Education. Generalizations of the result of this study should be made to the teachers who are in the same population of the study. Generalizations to other groups of teachers might be made only if the reader is willing to take responsibility for the validity of such generalizations.

Definition of Terms

Teacher is an elementary school teacher or secondary school teacher.

Elementary school teacher is a teacher who teaches in a public school in Bangkok which is under the Ministry of Education only, those who are under the Ministry of Interior are not included in this study.

Secondary school teacher is a teacher who teaches in a public school in Bangkok which is under the Ministry of Education.

Measurement need is defined as a lack of knowledge in a sub-area of tests and measurement as is disclosed by responses to groups of related test items that are designed to test desirable knowledge of measurement and evaluation.

Perceived need is defined as the feeling of lacking desirable knowledge in measurement and evaluation as disclosed by responses to questionnaire items.

Experience in teaching is defined as the number of years a teacher has been teaching in school. The higher the number of years a teacher has been teaching in school the more experience in teaching the teacher has. In this study, the teachers are classified into three groups: less than or equal to three years experience in teaching, between four to ten years experience in teaching, and more than ten years experience in teaching.

Certificate teacher is a teacher who had studied at a teacher training institution at least two years but not more than four years after grade 10, or its equivalence.

Degree teacher is a teacher who got at least a bachelor's degree in education, or its equivalence.

Item difficulty is defined as the percent of people who give an incorrect answer to the item.

Criterion score is the ideal mean. It is a point midway between the maximum possible score and the expected chance score (for example, ideal mean of 52 true-false test items = $1/2(52 + 52/2) = 39$). If the teachers have competence in measurement, the group mean should be higher than or equal to the ideal mean.

Overview

This study is reported in five chapters, followed by appendices.

In Chapter I, the introduction, need for the study, purpose of the study, statement of the problem, limitations of the study, and definition of terms used in this study were presented.

In Chapter II, the literature relevant to the general problem and related areas is reviewed. Description of the population, sampling procedure, the instrumentation, the design of the study and methods of analysis are discussed in Chapter III.

Chapter IV contains research data and results of the study. The final chapter contains a summary of the study, the conclusions and implications, and recommendations for further study.

CHAPTER II

REVIEW OF THE LITERATURE

The intent of this chapter is to review studies that are related to the problem as described in Chapter I. The review is divided into four sections: the role of teacher-made tests; the role of the teacher in testing and evaluation; the qualifications of the teacher as an evaluator; and improving the competence of teachers in measurement and evaluation.

The Role of Teacher-Made Tests

Classroom teachers are constantly searching for ways to improve their service to students. In line with this objective, they must have some procedures whereby they can reliably and validly evaluate how effectively their students have been taught. The classroom achievement test is one such tool.

Classroom tests may vary in form according to individual teachers' preference. Some teachers tend to favor the unstructured type in which the students must create the answers, as in the short answer or essay type. Other teachers prefer to use one or more of the structured or objective type tests in which the students select rather than create the answers. Still others use a combination of both unstructured and structured formats.

Most classroom tests must be prepared by the teacher who is teaching the class. While there are many standardized achievement tests available for broad areas of subject matter, few are specifically appropriate to the content and objectives of a unit of study. Teacher-made tests are better in the sense that they are more relevant to a teacher's particular objectives. As Mehrens and Lehmann (1978, p. 161) say, "Not only is the classroom teacher able to tailor the test to fit his particular objectives, but he can also make it fit the class and, if he wishes, fit the individual pupils. Commercially prepared tests, because they are prepared for use in many different school systems with many different curricular and instructional emphases, are unable to do these things as well as the teacher-made test."

Clearly, no standardized achievement test can completely serve the needs and purposes of every local situation (Noll, Scannell, and Craig, 1979, p. 148). Teachers usually feel that these tests do not adequately measure their own or the local objectives of instruction. Mehrens and Lehmann (1978, p. 161) have also pointed out that, "Commercially prepared achievement tests could be used to obtain some of the information needed by the teacher, and they could be used to motivate students. But, even in those schools that use commercial tests, it is unusual for such tests to be administered more than once a year. Also, the content of commercially prepared tests tends to lag, by a few years at least, recent curricular developments. Teacher-made tests are more likely to reflect today's curriculum. This is especially true in subject-matter areas such as science and social studies, which change rather rapidly in contrast to composition or literature."

Good classroom tests provide an efficient means for determining pupil ability and achievement. Ebel (1979, pp. 22-23) states, "The major function of a classroom test is to measure student achievement and thus to contribute to the evaluation of educational progress and attainments. A second major function of classroom tests is to motivate and direct student learning. The experience of almost all students and teachers supports the view that students do tend to study harder when they expect an examination than when they do not and that they emphasize in study those things on which they expect to be tested. Classroom tests have other useful educational functions. Constructing them, if the job is approached carefully, should cause an instructor to think carefully about the goals of instruction in a course."

It can be seen that a classroom test can serve many purposes, but it cannot do so with equal effectiveness. Mehrens and Lehmann (1978, p. 170) state that classroom achievement tests serve a variety of purposes, such as:

1) judging the pupils' mastery of certain essential skills and knowledge,
2) measuring growth over time,
3) ranking pupils in terms of their achievement of particular instructional objectives,
4) diagnosing pupil difficulties,
5) evaluating the teacher's instructional method,
6) ascertaining the effectiveness of the curriculum,
7) encouraging good study habits such as frequent review, and
8) motivating students.

Teacher-made tests are used principally for instructional functions (Stanley and Hopkins, 1972, p. 7). They provide a means of feedback to the teacher. Feedback from tests helps the teacher provide more appropriate instructional guidance for individual students as well as for the class as a whole. Well-designed tests may also be of value for pupil self-diagnosis, since they help students identify areas of specific weakness. The well-constructed tests can also motivate learning. As Mehrens (1979, p. 17) states, "Tests are credible instruments and will help motivate students and teachers." In general, students pursue mastery of objectives more diligently if they expect to be evaluated. The well-constructed examinations can give students an opportunity to test out their knowledge and constructive feedback can motivate students to improve on their performance.

In a study of classroom testing procedures and their influence on achievement, Marso (1970) found that:

1. unit testing does influence student achievement
2. feedback, pacing of learning, motivation and anxiety are related to student learning
3. testing procedures should incorporate frequent, graded, unit tests followed by class discussion
4. students with measured high test anxiety are not helped by frequent, graded, unit examinations with feedback.

Another study of the effect of frequent use of tests and feedback of test results was conducted by Feldhusen. Feldhusen (1964) reported that the majority of research reports prior to the time he undertook his study indicated that frequent use of teacher-made tests and feedback

from them resulted in better achievement and increased understanding of the concepts presented by the teacher. In his study, fifty-five college students in an introductory psychology class were given fourteen weekly quizzes consisting of 10 to 20 items. The quizzes were graded, returned, and on ten of the fourteen times when tests were given, classroom discussion ensued. At the end of the course, students were asked to respond anonymously to a questionnaire. He found that students consistently reported greater study and learning with periodic testing, and the anticipation of a forthcoming test may also affect students' "intention to remember" instructional content.

It is generally agreed that teacher-made tests are used because they enable the teacher to engage in continuous appraisal. There are, however, several limitations to the use of such tests that must be recognized if the teacher wants to make effective and efficient use of this means of measurement. As Schwartz and Tiedeman (1962, p. 110) say, "The foremost limitation of the use of teacher-made tests is the inadequate knowledge of most teachers concerning the principles of test construction. Test construction is a skill that can be learned but it takes time and practice to learn the skills of test construction. Another limitation that needs to be recognized is that good test items may take considerable time to prepare. Since there are limits to the time that is available to teachers, the problem of finding time to construct items may pose difficulties for some teachers. If, however, the teacher constructs test items on a daily basis, this limitation can be overcome."

Furthermore, Ebel (1979, p. 27) says that paper-and-pencil tests are well adapted to testing verbal knowledge and understanding and ability to solve verbal and numerical problems. These are important educational outcomes, but they are not all. One would not expect to get far using a paper-and-pencil test to measure children's physical development. Both performance tests of physical development and controlled observations of behavior in social situations would be expected to offer more promise than a paper-and-pencil test.

Constructing a satisfactory test is one of the hardest jobs a teacher has to perform. The process of constructing a good test item is deliberate and time-consuming; it demands an understanding of the objectives being assessed and of the examinees and their test-taking behavior. Teacher-made tests are of value only if they yield information that is used to improve the total teaching-learning process.

The Role of the Teacher in Testing and Evaluation

Teachers should be concerned with the types and levels of learning included in their courses from two perspectives: (1) the development and teaching of their courses, and (2) the assessment of their students' achievement (Lindvall, 1967, p. 3; Erickson and Wentling, 1976, p. 55). The essential purpose of teaching is to provide changes in students. Any program of instruction must be based upon and be guided by information concerning student aptitude, interest, and achievement. The classroom teacher should be guided by continuous information about student aptitude, interest and progress. Although the classroom teacher employs as much informal observation as possible as a means of acquiring

information about his students, it is necessary to use more formal procedures such as testing and other evaluation techniques.

Classroom realities compel teachers both to measure and evaluate student behaviors. The need to employ all applicable techniques of appraisal in the school has become a practical matter faced daily by teachers. As Goslin (1967, pp. 5-6) says, "The teacher occupies a central role in the testing and evaluation process for a number of reasons. First, the teacher is the primary point of contact between the child and the educational system, and what teachers say and do are major influences in the process whereby the child learns to assess his own abilities. Second, the teacher very often serves as the administrator and scorer of standardized tests, especially at the elementary level where testing specialists tend to be scarce. Even in situations where teachers are not directly involved in administering standardized tests, virtually all schools give teachers access to test scores. Finally, in a very real sense the teacher himself is being evaluated as a consequence of the performance of his pupils on standardized achievement tests. Teachers, therefore, are not disinterested observers of the testing process and may be expected to make efforts to improve the performance of their pupils on standardized tests, wherever this is practical. This, in turn, results in tests having a potential impact on school curricula insofar as what is taught and how it is taught is left to the teacher."

The teacher has a responsibility for appraising the individual differences among students in their achievement of various educational objectives (Thorndike and Hagen, 1969, p. 33). He must pass on to the next teacher a report of these differences, either in the form of a

mark or a specific recommendation, if the school is to provide an optimum learning environment for each child. Decisions about permitting students to pursue certain courses of study in high school, about admitting students to college, and about selecting students for certain occupations depend very largely upon judgments recorded by previous teachers concerning the competence of each student. The information on which these judgments are based is provided in considerable measure by tests.

Teachers must know how to perform certain aspects of measurement and evaluation themselves, such as constructing tests, giving grades, assessing potentialities, and interpreting standardized aptitude and achievement tests (Ebel, 1961a, pp. 19-32; Stanley, 1964, p. 5). They should know how to select from the many available tests, inventories, questionnaires, rating scales, checklists, and the like, those most suitable for a particular purpose. Besides being able to understand directions for administering, scoring, and interpreting tests, teachers should possess the higher ability to compare the most promising ones before the choice itself is made. This requires attaining various concepts necessary to understand test publishers' literature, reviews, and articles reporting test research.

Teachers sometimes dislike to assume the role of examiners (Ebel, 1979, p. 28). They, also, may be prone initially to frustration and disappointment when writing test items, possibly more so with one item format than another (Mehrens and Lehmann, 1978, p. 187). Ebel (1975) found that teachers did better in writing multiple choice test items than in writing true-false test items. He, however, comments that this

is hardly a fair comparison, since true-false items can be written more quickly by teachers, and responded to more quickly by students, than multiple-choice test items. Hence he feels that there is reason to question the recommendation that classroom teachers should generally give preference to multiple-choice over true-false test items.

The Qualifications of the Teacher as an Evaluator

The evaluation device used most frequently by the majority of teachers is undoubtedly the teacher-made test; therefore, it is essential that the beginning teacher be skilled in the development and use of such devices. Fortunately, the matter of how to develop and use classroom tests has received considerable attention in past years, and many useful criteria and suggestions have been prepared. The teacher who makes conscientious use of what is available in the area can greatly improve the quality of his tests (Lindvall, 1967, p. 30).

Many teachers admit that their tests do not adequately reflect the really important outcomes of their courses. Some are convinced that no test, and certainly no objective test, could adequately measure student achievement of their objectives (Ebel, 1972, p. 121). Sometimes teachers are embarrassed when they think of the way they judge their pupils. This is especially true after a parent-teacher conference if the teacher has not been able to explain the pupil's progress very effectively on the basis of measurement data collected (Lien, 1971, p. 20).

Schwartz and Tiedeman (1962) point out that one of the biggest errors teachers make in test planning is their tendency to wait until

shortly before an examination is scheduled to begin to write the items to be included in the test. Often the press of other duties seems much more important so that actual item writing is put off until the last minute. The result usually is that too many of the test items are poorly thought out, contain ambiguous terms, and in all too many cases, involve petty details instead of the more important and pervasive outcomes of learning. Ebel (1979, pp. 64-65) discusses some of the mistakes that teachers make in measuring educational achievement:

> First, they tend to rely too much on their own subjective judgment, and on unverified inferences.
>
> Second, some teachers feel obliged to use absolute standards in judging educational achievement, which can almost always be judged more fairly and consistently in relative terms. If most of the students in a class get A's on one test and most of the same students fail another, some teachers prefer to blame the students rather than the test.
>
> Third, teachers tend to put off test preparation to the last minute. A last-minute test is likely to be a poor test.
>
> Fourth, many teachers use tests that are too inefficient and too short to sample adequately the whole area of understanding and abilities that the course has attempted to develop.
>
> Fifth, teachers often overemphasize trivial details in their tests, to the neglect of understanding of basic principles and ability to make practical applications.
>
> Sixth, the questions that teachers write, both essay and objective, often suffer from lowered effectiveness due to unintentional ambiguity in the wording of the question or to inclusion of irrelevant clues to the correct response.
>
> Seventh, the inevitable fact that test scores are affected by the questions or tasks included in them tends to be ignored, and the magnitude of the resulting errors (called sampling errors) tend to be underestimated by those who make and use classroom tests.
>
> Finally, many teachers do not use the relatively simple techniques of statistical analysis to check on the effectiveness of their tests.

Are today's teachers being adequately prepared for performance of their evaluation responsibilities? Mayo (1967) found that graduating seniors in 86 teacher-training institutions did not demonstrate a very high level of measurement competence. Goslin (1967), in a study of the social consequences of testing and development of talent, found that about 60 percent of all teachers had only minimal exposure to training in test and measurement techniques. The unsatisfactory quality of the majority of teacher-made tests no doubt reflects this inadequacy in training. Not surprisingly, Goslin also found that teachers who had little preparation in tests and measurement tended to make little use of the pupil information obtained from standardized tests. Goslin (1967, p. 140) also states that "The role of teachers in testing is too important to be left to chance." Unfortunately, in view of studies by Mayo (1967) and Goslin (1967), Conant's recommendation (1963, p. 171) that instruction in tests and measurements be one of the essentials in teacher-training programs appears not to have been implemented adequately at many institutions.

Fleming (1971) claimed that not many teachers come to the classroom prepared to observe systematically, construct their own classroom tests, or to interpret the results of standardized tests regardless of the mode in which the scores are reported. She held that, "Part of pupil difficulties in the school have not only been due to inaccurate decisions by teachers but to the fact that teachers have been unskilled in constructing instructional cycles of relevant learning experiences based upon valid, definable goals." (Fleming, 1971, p. 71).

Roeder (1972) surveyed the qualifications or skill of the teacher as an evaluator. The 940 elementary teacher training institutions located in every state and the District of Columbia, were mailed a one-page questionnaire. The data indicate that 57.7 percent of the institutions which were surveyed, or 496 institutions, did not require their prospective elementary teachers to complete a course in evaluation; 12.1 percent (104) required nothing more than a one or two semester hour course; 17.8 percent (158) required a three semester hour course and only 1.4 percent (12) required four or more semester hours of course work in evaluation. Sixty-two institutions (7.2 percent) reported that instruction in evaluation was a component of another course, e.g., educational psychology. The data also indicate that in 1970, the vast majority of teachers who were graduated from accredited teachers colleges and awarded state certification, appeared to be better prepared to conduct an impromptu art lesson than they were to conduct, select, administer, score and interpret standardized and informal tests. Roeder concludes that it appears that even at institutions which do require a course in evaluation, the majority of teachers receive only a minimal exposure to the complex world of evaluation. Therefore, most of today's elementary teachers are not prepared to use tests.

Sor Wasna Pravalpruk (1974) studied the "Comparison among teachers in Khon Kaen, Thailand, to determine their testing needs." A Likert type questionnaire was constructed to measure the perceived needs, and the actual needs were measured by a random sampling of test items measuring knowledge on educational measurement. The questionnaire and test had twenty-eight items and twenty items respectively. They

were sent together to 400 teachers who were randomly selected from Khon Kaen province. Pravalpruk found that in the past, the in-service training program emphasized item editing in an attempt to improve the quality of the teacher-made test. The results of this training were reflected in the lower needs on both the actual and perceived needs in item editing than were found in other subject matter. The results of this study seem to indicate that it would be appropriate to emphasize item analysis procedures in future in-service programs because it was the area with the highest perceived needs score and was next to the highest in the test of actual needs. She also found that, in the item editing subject matter, teachers in the higher grades had less actual need than those who taught in lower grades.

In 1977, the Office of the National Education Commission (ONEC) studied the measurement competencies of the primary school teachers in Thailand. It was found that the third-grade teachers preferred to write multiple choice items but they did not use a table of specifications as the blueprint for test construction.

Yeh (1978), in a study of teacher use of test results, reported that only 50 percent of the teachers sampled were able to correctly interpret two standard scores commonly used in reporting standardized achievement results (percentile ranks and grade equivalents). She concluded from this that teachers need more knowledge about measurement.

> Given these findings about teachers' knowledge and the fact that teachers indicated they wanted more training on how to use and construct criterion-referenced tests, it may be that teachers need more training before any potential value of the test is realized. (Yeh, 1978, p. 42)

While there appears to be general agreement that teachers are not overly confident of their ability to interpret standardized test scores, the degree of confidence reported varies from researcher to researcher. Olejnik (1979), in a study conducted among non-test specialists (counselors, teachers and principals), found that over 90 percent of elementary and middle school educators indicated that they were at least "somewhat" confident of their ability to interpret test scores. The least confident were high school educationists. But when a mini-test similar to one given in college-level measurement courses was administered to the respondents, this self-reported "confidence" was not borne out. Most educationists correctly answered an item dealing with a percentile score (73%), yet a similar proportion missed an item that related norms to standards. They showed little understanding of the significance of stanine differences, and very few could properly interpret a grade equivalent score (12%). On the basis of his study, Olejnik concluded that in spite of self-reported confidence it appeared that non-measurement specialists needed additional assistance in the interpretation of standard scores.

A market survey of Stanford Achievement Test users was conducted by Stetz in 1977 (in Rudman et al., 1980). This study was aimed at determining the extent to which teachers and other educationists understand and accept standardized test results. He found that both teachers and administrators preferred grade equivalents and percentile ranks for meeting their assessment needs; 59% of the teachers surveyed chose these two scores for individual student evaluation, 56% chose these two scores for class evaluation purposes, 65% chose grade equivalents and

percentile ranks for measuring growth, and 67% preferred these two scores for reporting test results to parents. One would like to assume from this that those who showed such a strong preference for these two standard scores understood what they signified, but Olejnik's study (1979) does give one some pause.

The authorities seem to agree that testing is an integral part of teaching. Many teachers testify that the improvement of their skills in test construction has resulted in the improvement of their teaching. To make an appropriate and effective achievement test, one must have adequate knowledge of subject matter and skill in the techniques of test construction. Ebel (1961b, p. 68) has outlined six requisites for a teacher to be competent in educational measurement:

1) Know the educational uses, as well as the limitations, of educational tests.

2) Know the criteria by which the quality of a test should be judged and how to secure evidence relating to these criteria.

3) Know how to plan a test and write the questions to be included in it.

4) Know how to select a standardized test that will be effective in a particular situation.

5) Know how to administer a test properly, efficiently, and fairly.

6) Know how to interpret test scores correctly and fully, but with recognition of their limitations.

Sack (1979) studied the "Measurement competencies of educators defined through task analysis and differentiated by teaching area, grade level, and vocation." The principals at 292 randomly chosen northern Illinois public schools were asked to select a competent classroom teacher and a qualified staff member or administrator to anonymously

complete and return a task analysis questionnaire on measurement competencies. Educators and their responses were grouped by teaching area, grade level, and vocation with a view to specific measurement competencies preferred by categories of interest. A set of measurement competencies of acknowledged utility across nearly all educator categories tested was developed. These are:

1. Knowledge of advantages and disadvantages of standardized tests.
2. Understanding of the importance of adhering strictly to the directions and stated time limits of standardized tests.
3. Knowledge of general uses of tests, such as motivating, emphasizing important teaching objectives in the minds of pupils, providing practice in skill, and guiding learning.
4. Ability to state measurable educational objectives.
5. Knowledge of the techniques of administering a test.
6. Knowledge of effective procedures in reporting to parents.
7. Ability to interpret diagnostic test results so as to evaluate pupil progress.
8. Knowledge of limitations of tests that require reading comprehension.
9. Knowledge of limitations in interpreting IQ scores.
10. Understanding of the fact that interpretations of achievement from norms is affected by ability level, cultural background and curricular factors.

Improving the Competence of Teachers in Measurement and Evaluation

Measurement and evaluation are a part of every teacher's responsibilities. He must appraise the status and progress of the learner and make reports. A teacher can hardly be of maximum effectiveness today without knowing at least how to interpret and use the results of

standardized tests of readiness, intelligence, and achievement. In addition, the teacher must know how to measure and evaluate with instruments of his own devising. Since these are necessary, some instruction in the fundamentals of measurement should be included in the preparation of every teacher. Workshops, field courses, supervisory assistance, teachers' meetings, and professional reading are all helpful in improving teachers' skills in measurement and evaluation.

Improving classroom teacher competence in measurement and evaluation must not be treated as an isolated question, Margaret Stevenson (1959) says. It must be viewed in context, against the background of the numerous problems involved in establishing an adequate, democratically structured testing program in the school curriculum. While a great variety of things must be done to foster improvements in teacher competence in measurement, Ebel (1961b) suggests that special emphasis may be focused on only three:

1) increased attention to educational measurement in teacher-training programs;
2) provision of special testing services to teachers in school systems. This requires a school system to employ a staff member with special competence in testing; and
3) special organization of in-service training programs in measurement for teachers.

Noll (1961) recommends three possible ways for improving the preparation of teachers in measurement. The first would be to make a commitment to the policy of including a course in measurement as part of the requirement for a teacher's license or certificate. The second

would be to work for the strengthening of existing programs for preparation of teachers in all feasible ways in the area of measurement and evaluation but without a specific course requirement. Noll says that these are not necessarily antithetical or mutually exclusive but it seems likely that the requirement for a course in measurement of all prospective teachers would have the effect of reducing the emphasis on this topic in other courses where it is now usually included. A third possibility is the requirement of demonstrated proficiency in the area of measurement and evaluation on an examination, probably of a comprehensive objective nature.

In addition, Goslin (1967) states that school systems and testing specialists might be encouraged to initiate formal and informal training programs in measurement and evaluation for teachers.

The importance of in-service training programs in measurement has been recognized over the past several years. These programs, variously referred to as conferences, seminars, institutes, or workshops, have ranged from an afternoon lecture to a three-day preschool program with several follow-up meetings later in the year. Some of these programs were sponsored by a single school system and involved the teachers of that school in all subject areas and at all levels. Others reflected the interest of a single professional group, such as engineers or nurses (Ebel, 1961b).

A statewide research study in Tennessee about in-service education was conducted by Brimm and Tollett (1974). The purposes of this study were to identify the types of in-service education programs currently in use throughout the state and to ascertain teacher attitudes toward

in-service education programs. The results of the study can be summarized as the following:

1) The primary purpose of in-service programs is to upgrade the teacher's classroom performance.

2) The teacher should have the opportunity to select the kind of in-service activities which he feels will strengthen his professional competence.

3) In-service programs must include activities which allow for the different interests which exist among individual teachers. If teachers' professional growth is to be taken seriously, public school administrators and teachers must pool their knowledge and resources and seek to make in-service education more responsive to the needs and interests of practicing classroom teachers.

There seem to be some problems in in-service training in measurement and evaluation. Durost (1959) summarized the problems in in-service training in measurement as the following:

1) There are not enough leaders being trained in this area.

2) There is confusion and competition between professional groups training workers in the field of guidance, school psychology, and measurement per se as to who should be the person in the community with top responsibility for measurement.

3) Centralized training at the university level, no matter how good, will never diminish to the vanishing point the need for local in-service training at the community level because of unique local problems.

4) Teachers in general are afraid of measurement courses or even workshops in measurement because they are afraid of arithmetic, mathematics, statistics, etc.

5) Some teachers feel that the testing program can not genuinely help them to improve instruction.

Ebel (1961b) sees the two main weaknesses of the in-service training programs in measurement. Those are:

1) They are too brief. While an hour or two a year spent in considering measurement problems under the guidance of a specialist is far better than nothing at all, it is unreasonable

to suppose that satisfying enduring progress in solving the manifold problems of educational measurement, or in developing the requisite knowledge, understanding, and skills, can be made in so short a time.

2) They involve too much talking and too little doing. For the cultivation of a practical art like educational measurement, sound pedagogy requires a mingling of theory and practice.

The competence of the teacher in measurement can hardly be improved if the in-service program is not effective (Miller, 1977). Durost (1959) suggests specific steps to improve the in-service training program at the local level:

1) Preparation of local bulletins. A series of bulletins, supplementing the published materials concerning the tests in use in the county have been written tying in the testing program with the local program of instruction.

2) Use of school test coordinators. At the elementary level this may be the school principal or it may be a teacher with an interest in measurement who has been designated for this responsibility. These test coordinators meet regularly, especially before and after a scheduled testing program, to discuss problems involved in administering, scoring and interpreting the test results.

3) Extension courses in the area of measurement.

4) Faculty workshops. A considerable number of faculty workshops, varying in length from one to four or five sessions, have been held. These workshops concern themselves with the aspects of the total measurement problem which are important to the faculty at that moment.

5) Development of community interest in the measurement program, through a judicious use of local newspaper publicity, talk to parent-teacher associations, etc.

6) Use of local norms.

7) Provision of adequate physical facilities.

8) Use of demonstrations, lectures, etc., in the In-Service Training Center.

9) TV workshop on testing.

Ebel (1961b) proposes the ideal program of in-service training in measurement as the following:

> Suppose that a school administrator and his staff have decided to focus attention for a year on the improvement of classroom testing. Suppose they engage a specialist in educational testing to meet with them five times during the year, at intervals of six weeks or so, for a day or two. Participation in the initial program might well be limited to five, six, or seven groups of four to six teachers each.
>
> The goal of each group would be to make, to use, and to analyze a quality test in a subject which all members of the particular group were teaching. Examples of subject areas in which these tests might be developed are: fourth-grade mathematics; sixth-grade geography; eighth-grade English; or high-school history, chemistry, or economics.
>
> The first meeting of each participating group would be devoted to a description of the entire project, with special consideration of the first step - the preparation of specifications for the test to be developed. Sample specifications would be presented for study and analysis. Between the first and second sessions each teacher group would work out the specifications for its test. These could be reviewed at the second meeting, and work on item writing would be launched. The third meeting could be devoted to item review and test assembly, the fourth to test administration and analysis, and the fifth to a review of the test developed and of the entire project as a learning experience.

Ebel believes that a program like this would produce not only a handful of excellent tests but also a sizable group of teachers whose competence in measurement was vastly improved and, by current standard, highly respectable.

Zigarmi, Betz, and Jensen (1977) studied teachers' preferences in and perceptions of in-service education. They found that the in-service training programs will be useful if:

1) They are planned in response to the assessed needs of teachers and build on the interests and strengths of the teachers for whom they are designed.

2) They start with the assumption that teachers can be resources to each other and, therefore, these programs provide opportunities for teachers to share ideas and resources with each other.

Summary

Classroom teachers are constantly searching for ways to improve their service to children. In line with this objective, they are anxious to find new methods of measurement and evaluation and to enhance their skills in using these techniques (Ebel, 1961a; Stanley, 1964; Goslin, 1967; Lindvall, 1967; Erickson and Wentling, 1976).

The measurement of pupil achievement requires the extensive use of tests constructed by classroom teachers. This is so because many of the instructional outcomes can be measured by paper-and-pencil tests and because standardized tests are seldom well adapted to the particular objectives emphasized in teaching (Mehrens and Lehmann, 1978; Noll, Scannell, and Craig, 1979). In addition, teacher-made tests can be used for such a variety of instructional purposes (Schwartz and Tiedeman, 1962; Feldhusen, 1964; Marso, 1970; Stanley and Hopkins, 1972; Ebel, 1979; Mehrens, 1979). For example, the teacher may want to measure achievement at the end of a unit of work, diagnose a learning difficulty which has come to his attention, or check on how well the pupils have mastered a specific skill.

Constructing a good test is one of teachers' most difficult duties, and they, too, sometimes dislike to assume the role of examiners. The qualifications of teachers as the evaluators are widely reported as

less than satisfactory for most teachers. There is evidence that today's teachers are not being adequately trained for performance of their measurement and evaluation responsibilities (Conant, 1963; Mayo, 1967; Goslin, 1967; Fleming, 1971; Roeder, 1972; Pravalpruk, 1974; ONEC, 1977, Yeh, 1978; Olejnik, 1979; Sack, 1979). The results of these studies indicate that more publication and in-service programs need to be operated to improve the teacher's competencies.

There are a great variety of ways to improve the classroom teacher's competencies in measurement and evaluation such as: 1) an increase in attention to educational measurement in teacher-training programs, 2) employ a staff member with special competence in testing in the school system, 3) include a course in measurement as part of the requirement for a teacher's license or certificate, or 4) initiate formal and informal in-service training programs in measurement for teachers (Noll, 1961; Ebel, 1961b; Goslin, 1967).

The importance of in-service training programs in measurement has been recognized over several years. The primary purpose of these programs is to upgrade the teacher's classroom performance. To be effective, these programs must include activities which allow for the different interests which exist among individual teachers (Durost, 1959; Brimm and Tollett, 1974; Miller, 1977; Zigarmi, Betz, and Jensen, 1977).

CHAPTER III

PROCEDURES

This study can be classified as a comparative study. It was aimed at providing the data and information about the quality of the teacher as an evaluator to the administrators and educators in teacher training institutions in Thailand. Data were collected by questionnaires, sent through the mail or delivered personally. This chapter provides a description of the population, sampling procedure, instrument, data collection, and plan for data analysis.

The findings for the study are presented in Chapter IV and conclusions are given in Chapter V.

Population

Geographically, Thailand is in South-East Asia. The area of the country is 514,000 square-kilometers. The population of Thailand is about 46 million, of which 21% are students.

Bangkok, the capital of Thailand, has a population of four million, 25% of whom are students. There are 422 public elementary schools and 102 public secondary schools in Bangkok. Of the 422 elementary schools, 33 schools are under the Ministry of Education. The rest of them are under the Ministry of Interior. Ninty-two of 102

secondary schools are under the Ministry of Education. The other ten schools are demonstrative schools which are offered by universities in Bangkok.

The population of interest is teachers who have been teaching in public elementary and secondary schools, offered by the Ministry of Education in Bangkok. There are 1,599 teachers and 30,828 students in the 33 elementary schools, and 9,840 teachers and 203,476 students in the 92 secondary schools. The average ratio of teacher to students is 1:19 in elementary schools, and 1:21 for secondary schools.

The 33 elementary schools and 92 secondary schools are located in five different regions. The number of schools in each region is presented in Table 3.1. The breakdown of 11,439 teachers into six groups, according to level of school and level of teacher education, is presented in Table 3.2. The data from Table 3.2 show that 2% of 11,439

TABLE 3.1

Total Number of Schools in Bangkok Classified by Location, and Level of School

Location of School	Level of School	
	Elementary School	Secondary School
Region 1	9	23
Region 2	4	18
Region 3	7	15
Region 4	4	19
Region 5	9	17
Total	33	92

TABLE 3.2

Total Number of Teachers in 33 Elementary Schools and 92 Secondary Schools in Bangkok Classified by Level of School and Level of Teacher Education

Level of School	Level of Teacher Education			Total
	Master Degree	Bachelor Degree	Teaching Certificate	
Elementary School	21 (0.18%)	604 (5.28%)	974 (8.51%)	1,599 (13.98%)
Secondary School	255 (2.23%)	6,196 (54.17%)	3,389 (29.63%)	9,840 (86.02%)
Total	276 (2.41%)	6,800 (59.45%)	4,363 (38.14%)	11,439 (100%)

teachers hold a master's degree. Most of the elementary school teachers hold teaching certificates, but most of the secondary school teachers, in contrast, hold the baccalaureate degree. The percentage of the teachers who hold at least one baccalaureate degree for elementary school is 39% and 61% for secondary school.

Sampling Procedure

A list of all teachers in the 33 elementary schools and 92 secondary schools was provided by the Department of Elementary and Secondary Education, Ministry of Education. Since information concerning the elementary school teacher (such as level of teacher education, years of experience in teaching, etc.) was available for only 30 schools, the elementary school teachers were sampled from those 30 schools. In order to get the same number of schools in each region for both elementary and secondary schools, a number was assigned to every secondary school for the purpose of random selection. A table of random numbers was used, and a simple random sampling procedure was used to get a sample of 30 secondary schools (as shown in Table 3.3).

All of the teachers from 30 elementary schools and 30 secondary schools were classified into 12 groups, according to level of school, level of teacher education, and years of teaching experience. The number of teachers in each group is presented in Table 3.4. A random number was assigned to every teacher for the purpose of random selection. A systematic random sampling procedure was used to get a sample of 45 teachers in each of the 12 groups for the total of 540 teachers. Because of the difficulty of transportation, uncertainty of the mail,

and unwillingness of some teachers in answering the questionnaires, it was expected that the percentage of return would be less than a hundred percent. If less than a hundred percent of the questionnaires were returned in any of the 12 groups, the responses in each group were randomly selected again to get the same number of responses in each group. Because of the homogeneity of population, any bias contributed by the repeated sampling is minimal.

TABLE 3.3

Number of Elementary and Secondary Schools in Each Region Used in the Study

Location of School	Level of School	
	Elementary School	Secondary School
Region 1	9	9
Region 2	4	4
Region 3	7	7
Region 4	3	3
Region 5	7	7
Total	30	30

TABLE 3.4

Total Number of Teachers in 30 Elementary Schools and 30 Secondary Schools in Bangkok Classified by Level of School, Level of Teacher Education, and Years of Teaching Experience

Level of School	Level of Teacher Education	Years of Teaching Experience			Total
		0 - 3 Yrs.	4 - 10 Yrs.	> 10 Yrs.	
Elementary School Teacher	Certificate	105	129	638	872
	Degree	92	135	364	591
Secondary School Teacher	Certificate	435	289	379	1,103
	Degree	685	904	637	2,226
Total		1,317	1,457	2,018	4,792

Instrument

The pilot test contained 80 true-false items concerning basic knowledge in measurement and evaluation. These items were selected from two tests, 107 and 121 items each, by permission of Dr. Robert L. Ebel. These tests had been used for the final examination at least three times in a basic measurement course at Michigan State University (ED 465: Testing and Grading). Kuder Richardson Reliability # 20 of these two tests has varied between .88 to .91, mean item difficulty (percent of incorrect responses) has varied between .11 to .21, mean item discrimination (upper-lower difference) has varied between .18 to .27. The item selection was based on subject matter, item difficulty index, and item discrimination index.

The items in the pilot test covered basic knowledge in measurement and evaluation corresponding to four subject matters that teachers should know. The four areas are:

1. planning a classroom test,
2. item writing,
3. item analysis, and
4. test score statistics and marking system.

The pilot test was composed of 20 items in each subject matter area, with a total of 80 items. It contained 31 true statements and 49 false statements (see Appendix A). The test was then translated into Thai. The appropriateness of the test items and translation were checked by other Thai educators whose major area is measurement and evaluation.

The instrument was piloted out on forty Thai elementary school teachers and forty Thai secondary school teachers, for a total of eighty teachers. These teachers were omitted when selecting the sample. The test was distributed to this group in the first week of November, 1979, and all returns were collected in the first week of December, 1979. All responses were transferred to answer sheets and they were sent to the Scoring Office at Michigan State University for computing reliability and item analysis.

Most of the test items were difficult. The item difficulties (percent of incorrect responses) ranged from .06 to .89. Of the 80 items, 37 items (46%) had an index of difficulty above .60, only 11 items (14%) had the index of difficulty below .20. The discrimination indices ranged from -.19 to .47. The Kuder Richardson Reliability # 20 was .40, mean item difficulty was .53, and mean item discrimination was .15. The items for each subscale were selected separately for the final test. Within each of the four subject matters, the item discrimination indices and item difficulties were considered in deleting items. The final test contained fifty-two items, thirteen items for each subscale.

Among twenty items in Planning a Classroom Test, seven items (items 2, 7, 9, 12, 17, 18, and 19) had negative discrimination indices (ranged from -.19 to .00). Five of these seven items had high difficulty indices (ranged from .51 to .81), items 2 and 9 had difficulties of .45 and .06 respectively. Therefore, all seven items were deleted. Among the items in Item Writing, items 22, 26, 28, 33, 34, 38, and 40 were deleted because of low discrimination indices (ranged from -.05 to

.09) and high levels of difficulty. The item difficulties of the first six deleted items ranged from .56 to .89. Item 40 had the lowest difficulty index among these seven items, its difficulty index was .43.

In the Item Analysis group, the deleted items were numbers 42, 43, 46, 49, 55, 56, and 58 with the range of discrimination indices between -.05 to .09, and the range of the levels of difficulty between .06 to .90. Only two of them had difficulty indices below .20. In the last subject matter area, Test Score Statistics and Marking System, items 64, 67, 68, 72, 73, 74, and 75 were deleted. The difficulty indices of these seven items ranged from .21 to .83. Four of them had an index of difficulty above .71, only two items had an index of difficulty below .21. The discrimination indices of these seven items ranged from -.09 to .24.

The final instrument (see Appendix B) was composed of two parts:

1. A questionnaire concerning the general information of the teachers (3 items), opinion on the national testing program (1 item), and teachers' experiences in measurement and evaluation (4 items). This part contained eight items.

2. The second part was composed of 52 true-false items of which 20 statements were true and 32 statements were false. These test items measured the basic knowledge in measurement and evaluation. The mean item difficulty on the pilot test for the Planning a Classroom Test subscale was .50, for the Item Writing subscale it was .49, for the Item Analysis subscale it was .49, and it was .55 for the Test Score Statistics and Marking System subscale.

Data Collection

The instrument was sent to teachers in the sample in the first week of February, 1980, both by mail and personal delivery, school by school. To insure getting back a large majority of responses, the follow-up was done weekly either by letter or personal contact. Because of the personal contacts and some help from the school principal, sixty-nine percent of the responses (374 responses) were returned by the last week of March, 1980. It was noticed that the less personal relationship between the agent collecting data and the teachers in the sample, the less likely that the responses would be returned. Since the number of the returned responses in each group ranged between 30 to 33 (see Table 3.5), fourteen responses (3.7% of responses) were randomly thrown out in order to get thirty subjects in each group for the total of 360 subjects. Because of the homogeneity of population and only 3.7% of responses were randomly thrown out, it is believed that any distortion of data is small.

Design

The design of this study is a 2 x 2 x 3 factorial design with four repeated measures. The four areas of basic knowledge in measurement and evaluation were: planning a classroom test; item writing; item analysis; and test score statistics and marking system. Table 3.6 presents the layout of the three indpendent variables: level of school; level of teacher education; and teaching experience. The design is crossed and balanced with thirty observations per cell.

TABLE 3.5

Total Number of Returned Responses Classified by
Level of School, Teacher Education,
and Years of Teaching Experience

Level of School	Level of Teacher Education	Years of Teaching Experience			Total
		0 - 3 Yrs.	4 - 10 Yrs.	> 10 Yrs.	
Elementary School	Certificate	30	31	30	91
	Degree	32	31	32	95
Secondary School	Certificate	31	32	33	96
	Degree	32	30	30	92
Total		125	124	125	374

TABLE 3.6

Design of the Three Independent Variables:
Level of School, Level of Teacher
Education, and Teaching Experience

Level of School	Level of Teacher Education	Teaching Experience	Measurement			
			M_1	M_2	M_3	M_4
Elementary School	Certificate	0 - 3 yrs. n=30 4 - 10 yrs. n=30 > 10 yrs. n=30				
	Degree	0 - 3 yrs. n=30 4 - 10 yrs. n=30 > 10 yrs. n=30				
Secondary School	Certificate	0 - 3 yrs. n=30 4 - 10 yrs. n=30 > 10 yrs. n=30				
	Degree	0 - 3 yrs. n=30 4 - 10 yrs. n=30 > 10 yrs. n=30				

M_1 = planning a classroom test

M_2 = item writing

M_3 = item analysis

M_4 = test score statistics and marking system

The following null hypotheses were tested:

1. There is no difference in measurement needs between elementary school teachers and secondary school teachers in Bangkok.
2. There is no difference in measurement needs between teachers who have a higher level of education and those who have a lower level of education.
3. There is no difference in measurement needs between teachers who have more teaching experience and those who have less teaching experience.
4. There is no difference in the needs among the four subject matters in measurement covered in the study.
5. There is no level of school by level of teacher education interaction.
6. There is no level of school by teaching experience interaction.
7. There is no level of teacher education by teaching experience interaction.
8. There is no level of school by level of teacher education by teaching experience interaction.
9. There is no subject matter by level of school interaction.
10. There is no subject matter by level of teacher education interaction.
11. There is no subject matter by teaching experience interaction.
12. There is no subject matter by level of school by level of teacher education interaction.
13. There is no subject matter by level of school by teaching experience interaction.

14. There is no subject matter by level of teacher education by teaching experience interaction.
15. There is no subject matter by level of school by level of teacher education by teaching experience interaction.

Analysis

The univariate analysis and multivariate analysis of repeated measures were employed to test the research hypotheses. If a source of variation was found to be significant, the Tukey or Scheffe technique was used to test the differences between groups. Graphic presentation was considered when there was a significant interaction term.

In addition to testing these hypotheses of interest, the Z-test was employed to test the differences between group means and criterion scores. The analysis was done individually on both total score and subscale scores. Descriptive data are provided such as the mean and standard deviation for each cell in the design. Descriptive data on opinion on training program in measurement and evaluation, opinion on national testing program, and relationship between measurement needs and perceived needs in basic measurement are also provided.

Summary

A list of 30 elementary schools and 30 secondary schools and a list of all 4,792 teachers in Bangkok was provided by the Ministry of Education. The teachers were classified into twelve groups according to two categories of level of school, two categories of level of teacher education, and three categories of teaching experience. Forty-five teachers were randomly selected within each group for the total of 540

teachers. A questionnaire containing the attitude items and test items was tried out in November, 1979. All responses were transferred to answer sheets and they were sent to the Scoring Office at Michigan State University for computing reliability and item analyses.

The revised edition of the questionnaire was sent to 540 teachers in February, 1980. Since only sixty-nine percent of the responses were returned, fourteen responses were randomly thrown out in order to get thirty subjects in each group for the total of 360 subjects. All responses were coded and data punch cards were produced. Univariate analysis and multivariate analysis of repeated measures were used to test the hypotheses. All of the hypotheses were tested at the significant level of .05. Graphic presentation was considered when there was a significant interaction. The Tukey technique (for equal cell size) or the Scheffe technique (for unequal cell size) was used to test the differences between groups. Additional descriptive data were presented.

CHAPTER IV

ANALYSES AND RESULTS

The final instrument was sent to teachers in the sample in the first week of February, 1980. At the end of the second week of April, 1980, all data were coded. Responses to each test item were coded 0 and 1, with 0 assigned to incorrect responses and 1 assigned to a correct response. Coding systems for all descriptive data were also created and used to record the information. All data cards were punched, and were ready to use in May, 1980.

Two computer programs were used to analyze the data in this study. The Statistical Package for Social Sciences (SPSS, 1975) program was used to compute Cronbach's Alpha reliability and the item analyses. It was also used to perform the cross-tabulation or contingency table, cell means and cell standard deviations, analysis of variance, and Tukey and Scheffe ranges for significant main effects. The Multivariance program[1] was used to perform univariate analysis and multivariate repeated measures analysis. All computer programs were executed by the CDC-6500 at the Computer Center, Michigan State University.

---

[1]Jeremy D. Finn's Multivariance - Univariate and Multivariate Analysis of Variance, Covariance, and Regression. Modified and adapted for use on the CDC-6500 at Michigan State University by Verda M. Scheifley and William H. Schmidt, 1973.

Before any analysis was done, the instrument was re-evaluated. The Cronbach's Alpha reliability was .41. Mean item difficulty (percent of individuals giving an incorrect answer) was .49. Of 52 items, 21 items had indices of difficulty above .61, only eight items had indices of difficulty below .20. Twenty-nine of 52 items had discrimination indices (r-biserial correlation) between .21 to .40, four items had indices of discrimination above .41 and 19 items had indices of discrimination below .20.

In this chapter, the results of the data analyses are presented. The descriptive data are discussed first. The latter section deals with testing the hypotheses of interest and also additional analyses besides the hypotheses of interest.

Descriptive Data

Of all 360 teachers, 75% (269 teachers) had taken a college measurement course. Of these 269 teachers who had taken a college measurement course, 47% (168 teachers) hold at least a bachelor's degree, and 28% (101 teachers) hold a teaching certificate.

Eighty-five of 107 teachers who attended the training program took a college measurement course, 22 teachers did not take a course. Only 19% (69 teachers) neither took a college measurement course nor attended the training program in measurement and evaluation.

Ninety-six percent of those who attended the training program in measurement (107 teachers) felt the training program worthwhile, and 243 of 300 teachers (81%) want to participate in the training program in measurement if it is offered in the future.

The majority of teachers liked the idea of national testing. Sixty-three percent (226 teachers) responded yes to that question. Of these 226 teachers, 119 teachers (53%) are elementary school teachers, 107 teachers (47%) are secondary school teachers, 111 teachers (49%) hold a teaching certificate, and 115 teachers (51%) hold at least a bachelor's degree.

Repeated Measures Analysis on Measurement Scores

Summary data, cell means, and cell standard deviations of the four subscales of the measurement scores and total score are presented in Table 4.1. To get a general profile of the four dependent variables, the means and standard deviations for the entire sample are also presented. The maximum possible score for each subscale was 13. The highest mean was on the Item Analysis subscale (6.87), the next highest was on the Test Score Statistics and Marking System subscale with a mean of 6.73. Third was the Item Writing subscale with a mean of 6.66. The lowest mean was on the Planning a Classroom Test subscale (6.21). The cell standard deviations varied from 1.15 to 2.06.

The results of the multivariate repeated measures analysis are presented in Table 4.2. Fifteen hypotheses were tested. Four of them were on the main effects, the other eleven hypotheses concerned the interactions. Four of the fifteen null hypotheses were rejected at $\alpha = .05$. The following are the hypotheses which have been rejected:

- the main effect of teacher education ($F = 23.91$, $p < .0001$),
- the interaction between level of school and teacher education ($F = 5.0161$, $p < .0258$),

TABLE 4.1

Means and Standard Deviations of Four Subscales and Total Scores

School	Education	Experience	Measurements PL	WR	AN	ST	Total
Elementary School	Certificate	0 - 3 yrs.	6.03	6.63	6.77	6.20	25.63
		n = 30	1.71	1.92	1.36	1.37	3.50
		4 - 10 yrs.	6.03	6.47	7.33	6.70	26.53
		n = 30	1.99	1.80	1.69	1.78	3.76
		> 10 yrs.	5.60	6.40	6.37	6.27	26.63
		n = 30	1.54	1.99	1.59	1.64	3.94
	Degree	0 - 3 yrs.	6.00	6.67	7.10	7.40	27.17
		n = 30	1.74	1.95	1.58	1.73	3.91
		4 - 10 yrs.	6.27	6.47	6.80	6.40	25.93
		n = 30	1.60	1.74	1.47	2.06	4.35
		> 10 yrs.	6.00	6.80	6.67	7.57	27.03
		n = 30	1.84	1.75	1.24	1.36	3.68
Secondary School	Certificate	0 - 3 yrs.	5.83	6.37	6.70	6.40	25.30
		n = 30	1.44	1.77	2.02	1.61	4.41
		4 - 10 yrs.	5.80	6.27	6.67	6.37	25.10
		n = 30	1.88	1.64	1.63	1.73	3.82
		> 10 yrs.	6.07	6.43	6.87	6.13	25.50
		n = 30	1.57	1.59	1.63	1.83	3.86
	Degree	0 - 3 yrs.	6.60	7.30	7.50	7.13	28.53
		n = 30	1.57	1.70	1.89	1.53	4.25
		4 - 10 yrs.	6.97	7.37	6.83	6.90	28.07
		n = 30	1.83	1.77	1.15	1.65	3.27
		> 10 yrs.	7.30	6.80	6.80	7.37	28.27
		n = 30	1.82	1.73	1.63	1.79	4.73
Entire Sample			6.21	6.66	6.87	6.73	26.48
			1.76	1.79	1.59	1.73	4.12

PL = Planning a Classroom Test
WR = Item Writing
AN = Item Analysis
ST = Test Score Statistics and Marking System

TABLE 4.2

Multivariate Repeated Measures Analysis
on Basic Measurement Scores

Sources of Variation	DF	F	P Less Than
School (S)	1,348	2.3227	.1285
Education (E)	1,348	23.9135	.0001*
Teaching Experience (T)	2,348	.1960	.8222
SE	1,348	5.0161	.0258*
ST	2,348	.2536	.7762
ET	2,348	1.0875	.3383
SET	2,348	1.2252	.2950
Subject Matter (M)	3,346	11.8591	.0001*
MS	3,346	1.5126	.2110
ME	3,346	2.8938	.0354*
MT	6,692	.9080	.4885
MSE	3,346	1.0388	.3755
MST	6,692	1.0782	.3740
MET	6,692	1.6543	.1297
MSET	6,692	.4990	.8094

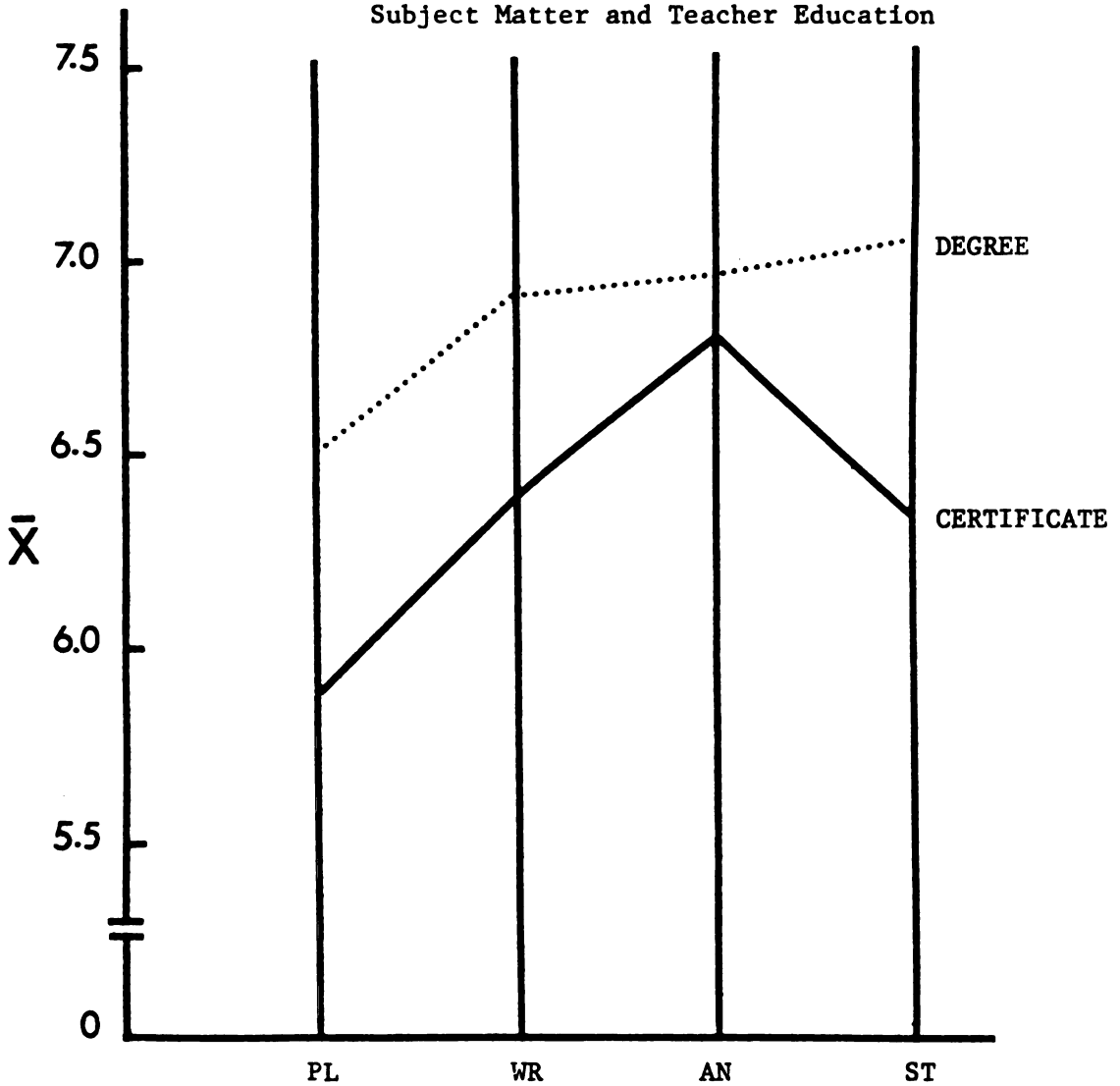
*The test is significant at $\alpha = .05$ level.

- the subject matter main effect ($F = 11.8591$, $p < .0001$),
- and, the subject matter by teacher education interaction ($F = 2.8938$, $p < .0354$).

Since the interaction between subject matter and teacher education was significant, the general profile was not interpreted. One profile was made for all certificate teachers and another for degree teachers. Figure 1 presents these two profiles. Means of each subscale of measurement for certificate teachers varied from 5.89 (for Planning a Classroom Test) to 6.78 (for Item Analysis). For the degree teachers, the means varied from 6.52 (for Planning a Classroom Test) to 7.13 (for Test Score Statistics and Marking System). The differences between certificate teachers and degree teachers on Planning a Classroom Test, Item Writing, Item Analysis, and Test Score Statistics and Marking System were -.63, -.47, -.17, and -.79 respectively. The interaction was ordinal with respect to teacher education. The largest difference between certificate teachers and degree teachers was found on the mean of Test Score Statistics and Marking System subscale, the smallest difference of these two groups was found on the Item Analysis subscale.

A multivariate analysis of variance was performed to test the significant differences between the mean of each subscale of certificate teachers and degree teachers. The multivariate F ratio was significant at $\alpha = .05$ ($F = 7.6698$, $p < .00001$), and the univariate F ratios indicated that there were significant differences in scores of the two groups of teachers on Planning a Classroom Test subscale, Item Writing subscale, and Test Score Statistics and Marking System subscale. A

Figure 1. Graph Presentation of Cell Means for Subject Matter and Teacher Education



PL = Planning a Classroom Test
WR = Item Writing
AN = Item Analysis
ST = Test Score Statistics and Marking System

	PL	WR	AN	ST
Certificate	5.89	6.43	6.78	6.34
Degree	6.52	6.90	6.95	7.13

significant difference between the two groups, however, was not found on the Item Analysis subscale (see Table 4.3).

TABLE 4.3

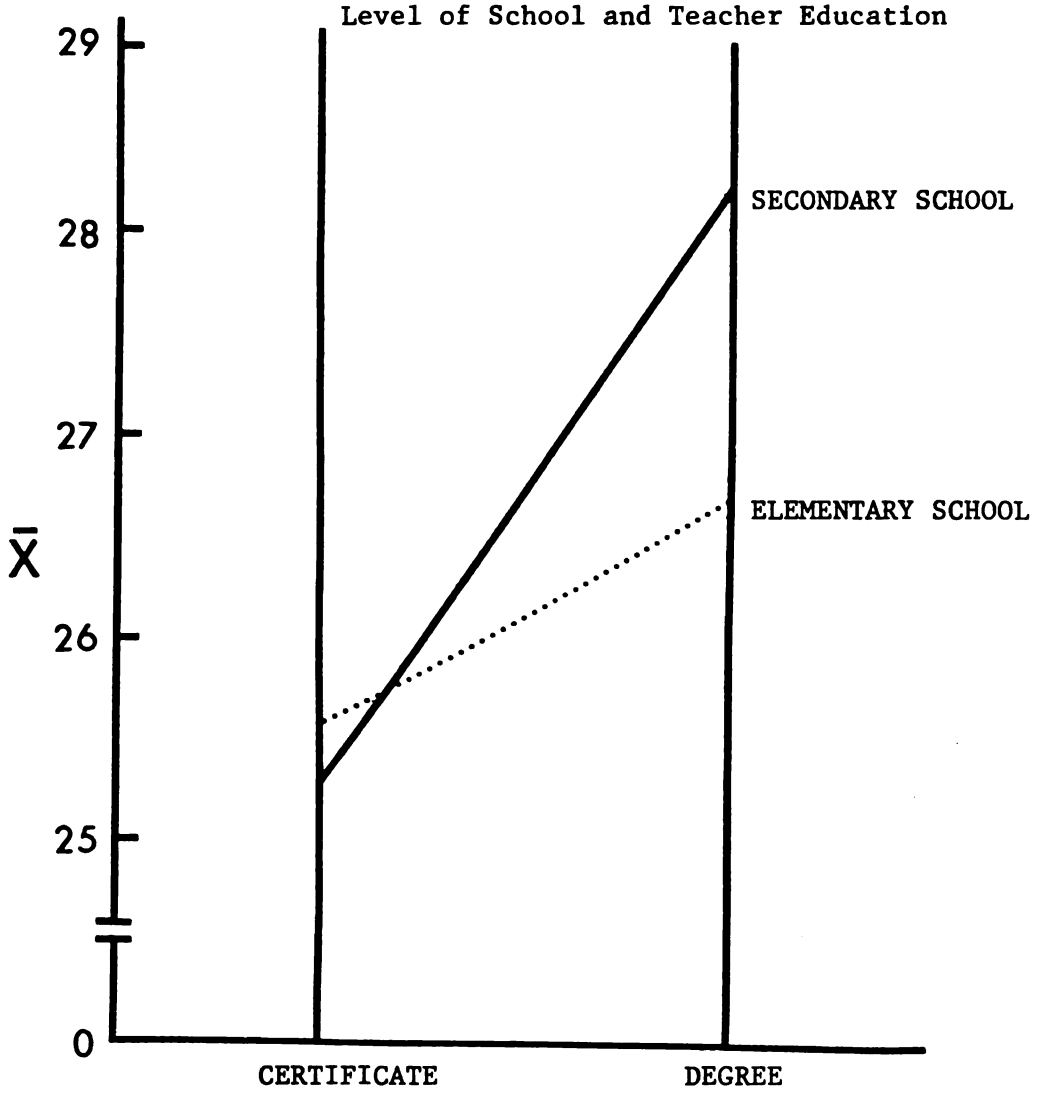
Univariate Analysis of Variance on Subscale Scores of Certificate Teachers and Degree Teachers

Variables	DF	F	Signif. of F
Plan	1,358	11.8022	.0007
Write	1,358	6.3681	.0121
Analy.	1,358	.9845	.3218
Stat.	1,358	19.3695	.0000

Plan = Planning a Classroom Test
Write = Item Writing
Analy. = Item Analysis
Stat. = Test Score Statistics and Marking System

The significant interaction between level of school and teacher education was also observed (Figure 2). The mean of certificate teachers in elementary school was slightly higher than the mean of certificate teachers in secondary school (25.60 and 25.30 respectively), but in contrast, the mean of degree elementary school teachers was lower than the mean of degree secondary school teachers (26.71 and 28.29 respectively). The mean of degree teachers in elementary school was higher than the mean of certificate teachers but it was not large enough to be significant at $\alpha = .05$ level. In secondary school, the difference between the mean of degree teachers and the mean of certificate teachers was significant at $\alpha = .05$ level.

Figure 2. Graph Presentation of Cell Means for Level of School and Teacher Education



	Certificate	Degree
Elementary School	25.60	26.71
Secondary School	25.30	28.29

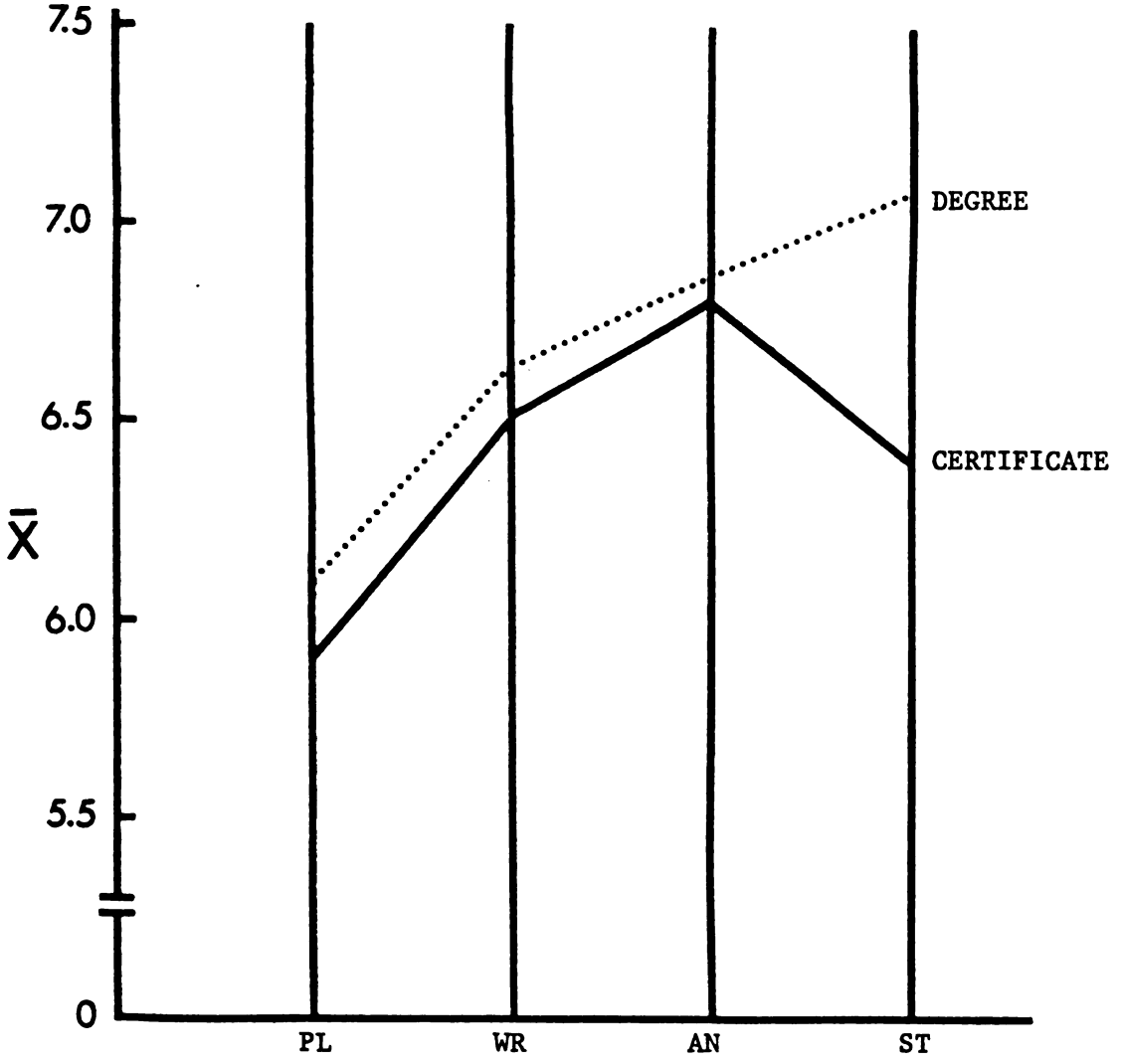
Since the interaction between level of school and teacher education was significant, the follow-up analyses were performed using multivariate analysis of variance testing the significant differences on subscale scores between degree teachers and certificate teachers in elementary school and in secondary school separately. Cell means for subject matter and level of education of elementary school teachers are presented in Figure 3, and in Figure 4 for secondary school teachers. The multivariate F ratio of secondary school teachers was significant at $\alpha = .05$ ($F = 7.4127$, $p < .00002$). The univariate F ratio indicated that there were significant differences among certificate teachers and degree teachers in secondary school on only three subscales; Planning a Classroom Test, Item Writing, and Test Score Statistics and Marking System (see Table 4.4). Means of degree teachers in elementary school were higher than means of certificate teachers for every subscale, but the amount of differences were not large enough to be significant at .05 level.

TABLE 4.4

Univariate Analysis of Variance on Subscale Scores of Certificate and Degree Secondary School Teachers

Variables	DF	F	Signif. of F
Plan	1,178	17.5710	.0000
Write	1,178	10.0430	.0018
Analy.	1,178	1.4391	.2319
Stat.	1,178	11.0521	.0011

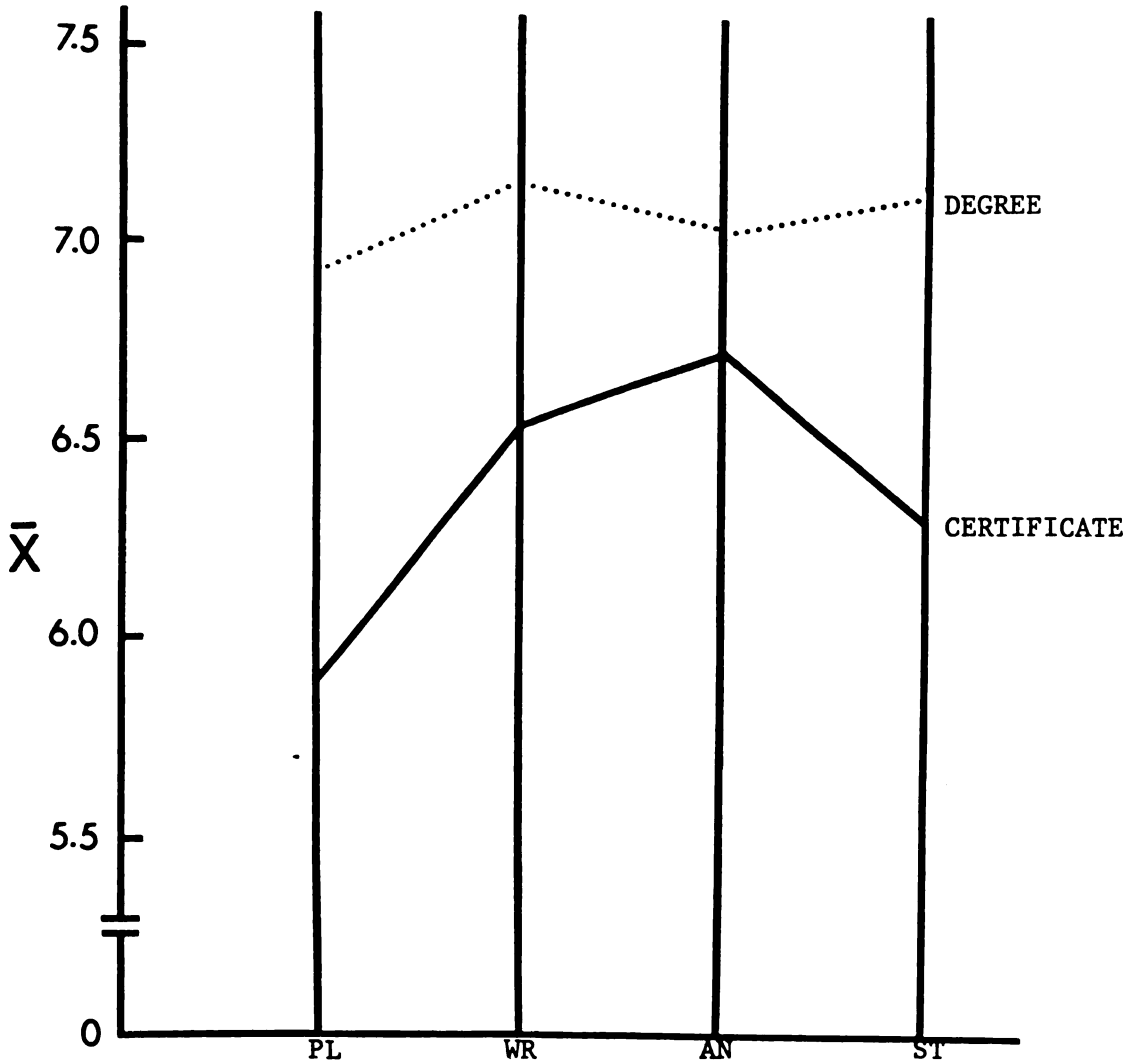
Figure 3. Graph Presentation of Cell Means for Subject Matter and Level of Education of Elementary School Teachers



PL = Planning a Classroom Test
WR = Item Writing
AN = Item Analysis
ST = Test Score Statistics and Marking System

	PL	WR	AN	ST
Certificate	5.89	6.50	6.82	6.39
Degree	6.09	6.64	6.86	7.12

Figure 4. Graph Presentation of Cell Means for Subject Matter and Level of Education of Secondary School Teachers



PL = Planning a Classroom Test
WR = Item Writing
AN = Item Analysis
ST = Test Score Statistics and Marking System

	PL	WR	AN	ST
Certificate	5.90	6.56	6.74	6.30
Degree	6.96	7.16	7.04	7.13

Additional Analyses

Besides testing the hypotheses of interest, further analyses were done to compare measurement needs (indicated by lower scores from the test) among various groups of teachers as defined by the following independent variables: taking a college measurement course, attending the training program in measurement, favoring a national testing program, and region of school. The analyses were also done to observe the relationship between perceived needs (indicated by the feeling of lacking desirable knowledge in measurement) and measurement needs, and to compare mean differences between group means and criterion scores (ideal mean). Group means, F ratios testing the difference between groups, and the probabilities of the F ratios are presented in Tables 4.5 to 4.10.

Multivariate analysis of variance was used to test the mean differences of all four subscale scores between teachers who took a college measurement course and those who did not take a course. The multivariate F ratios of the entire sample, of the degree teachers, and of the secondary school teachers were significant at $\alpha = .05$ level ($F = 5.3747$, $p < .0003$; $F = 3.0278$, $p < .0191$; $F = 4.4114$, $p < .0020$ respectively). The significant differences on subscale scores were not found in certificate teachers groups or in elementary school teachers groups.

Tables 4.5, 4.6, and 4.7 present the univariate analysis of variance on subscale score and total score of teachers who had and those who had not taken a college measurement course, the analyses were done individually for the entire sample, for degree teachers group, and for secondary school teachers group respectively. It was observed that the teachers who had taken a college measurement course received higher

TABLE 4.5

Analysis of Variance on Subscale Scores and Total Score of Teachers who Took a College Measurement Course and Those Who Did Not Take a Course

Variables	College Measurement		Univariate	
	Took $\underline{n}$=269 ($\bar{X}$)	Did Not Take $\underline{n}$=91 ($\bar{X}$)	F	P Less Than
Plan	6.29	5.98	2.0930	.1488
Write	6.83	6.19	8.8551	.0031
Analy.	6.96	6.58	3.9065	.0489
Stat.	6.93	6.18	13.1894	.0003
Total	27.00	24.92	18.1281	.0000

Plan = Planning a Classroom Test
Write = Item Writing
Analy. = Item Analysis
Stat. = Test Score Statistics and Marking System

total scores than those who had not taken a course in all three groups of teachers, the mean differences were 2.08, 3.75, and 2.75 respectively, the significant differences were found at $\alpha = .05$ level. On the four subscales, the teachers who had taken a college measurement course, in all three groups mentioned above, received higher scores on Item Writing and Item Analysis subscales than those who had not taken a course, the significant differences were found at $\alpha = .05$ level. No significant difference was found on Planning a Classroom Test subscale in all three groups of teachers. On Test Score Statistics and Marking System

subscale, the significant difference was found in the entire sample and in the secondary school teachers group at α .05 level in favor of those who had taken a college measurement course, but it was not found in the degree teachers group. Multivariate analysis of variance was also used to test the mean differences of all four subscales between the teachers who attended the training program in measurement and those who did not attend the program. No significant differences were found in any of the subscales. The total mean of the teachers who attended the training program was 26.4 and it was 26.5 for those who did not attend the program.

TABLE 4.6

Analysis of Variance on Subscale Scores and Total Score of Degree Teachers Who Took a College Measurement Course and Those Who Did Not Take a Course

Variables	College Measurement		Univariate	
	Took n=168 (X)	Did Not Take n=12 (X)	F	P Less Than
Plan	6.55	6.17	.5118	.4753
Write	7.00	5.50	8.2608	.0045
Analy.	7.01	6.08	4.2666	.0403
Stat.	7.19	6.25	3.3900	.0673
Total	27.75	24.00	9.8179	.0020

TABLE 4.7

Analysis of Variance on Subscale Scores and Total Score of Secondary School Teachers Who Took a College Measurement Course and Those Who Did Not Take a Course

Variables	College Measurement		Univariate	
	Took n=137 ($\bar{X}$)	Did Not Take n=43 ($\bar{X}$)	F	P Less Than
Plan	6.53	6.12	1.7659	.1856
Write	6.98	6.05	9.8969	.0019
Analy.	7.03	6.47	3.7478	.0545
Stat.	6.92	6.07	8.2370	.0046
Total	27.45	24.70	14.4437	.0002

The comparison between teachers who favored a national testing policy and those who did not favor the policy was performed by using multivariate analysis of variance. The teachers who favored a national testing policy received higher scores than those who did not favor the policy in every subscale (mean differences were .22, .22, .39, and .40 respectively) but the differences were not large enough to be significant at $\alpha = .05$ ($p < .0642$). A similar comparison was made between five groups of teachers, classified by the location of the school in which they were teaching. Slight differences between the means of those five groups of teachers were observed (the mean for Region 1 was 26.36,

for Region 2 it was 26.43, for Region 3 it was 25.86, for Region 4 it was 26.68, and for Region 5 it was 27.32) but the differences were not large enough to be significant at the .05 level ($p < .2473$).

Table 4.8 presents cell means for all four subscales of certificate and degree teachers who were classified into four groups according to the area of measurement they thought they knew most. Multivariate analysis of variance was used to determine if there was any significant difference between the means within four subgroups of certificate teachers and within four subgroups of degree teachers in each subscale of measurement. It was found that there were no significant differences either in certificate teacher groups or in degree teacher groups in any of the subscales. For example, it was observed that the teachers who thought they knew the most in Planning a Classroom Test did not get the highest score in this subscale when compared with the other three subscales. The same result occurred in the other three groups of teachers.

A similar comparison was made between four groups of teachers who were classified by the area of measurement they thought they knew least. Cell means and number of teachers in each group are presented in Table 4.9. No significant differences were found in any of the subscales. The data from Tables 4.8 and 4.9 indicated that there was no relationship between perceived needs (indicated by the feeling of lacking desirable knowledge in measurement) and measurement needs (indicated by lower test score).

The comparison between means of the entire sample and criterion scores (ideal means) were done by using the Z-test. The analyses were done individually on both total mean and subscale means. Means of the

entire sample, criterion scores, Z ratios testing the differences between group means and criterion scores, and the probabilities of the Z ratio are presented in Table 4.10. It was found that the total mean and the mean of each subscale were lower than the criterion scores, and the mean differences were significant at $\alpha = .05$ level.

TABLE 4.8

Presentation of Cell Means for All Four Subject Matter Areas of Certificate and Degree Teachers Who Were Classified into Four Different Groups According to Area of Measurement They Thought They Knew Most

	Specialized Area in Measurement	Test Score ($\bar{X}$)				N
		Plan	Write	Analy.	Stat.	
Cert. Teachers	Plan	5.72	6.38	6.38	6.54	39
	Write	5.86	6.37	6.89	6.46	87
	Analy.	6.22	6.86	6.89	5.94	36
	Stat.	5.78	5.94	6.89	6.17	18
Degree Teachers	Plan	6.41	6.94	7.00	7.41	34
	Write	6.05	6.59	6.73	6.92	75
	Analy.	7.03	7.15	7.06	7.06	33
	Stat.	7.11	7.26	7.24	7.34	38

TABLE 4.9

Presentation of Cell Means for All Four Subject Matter Areas of Certificate and Degree Teachers Who Were Classified into Four Different Groups According to Area of Measurement They Thought They Knew Least

	Weak Area in Measurement	Test Score ($\bar{X}$)				N
		Plan	Write	Analy.	Stat.	
Cert. Teachers	Plan	6.24	6.39	6.71	6.32	38
	Write	6.75	6.00	7.00	6.63	8
	Analy.	5.62	6.48	6.83	6.44	71
	Stat.	5.89	6.44	6.75	6.22	63
Degree Teachers	Plan	6.91	6.91	7.06	7.39	33
	Write	7.07	6.93	7.07	6.96	27
	Analy.	6.26	6.84	6.99	6.97	74
	Stat.	6.35	6.98	6.74	7.28	46

TABLE 4.10

Comparison Between Sample Means and Criterion Scores

Variables	$\bar{X}$	Criterion Score*	Z	P Less Than
Plan	6.21	9.75	-38.03	.0000
Write	6.66	9.75	-32.75	.0000
Analy.	6.87	9.75	-34.32	.0000
Stat.	6.74	9.75	-33.04	.0000
Total	26.48	39.00	-57.72	.0000

*Criterion score (ideal mean) is defined as a point midway between the maximum possible score and the expected chance score (for example, Criterion Score of 52 true-false test items = 1/2(52+52/2) = 39).

Summary

A descriptive discussion on information from the questionnaire items was presented first. Then the multivariate repeated measures analysis was employed to test the fifteen null hypotheses. The hypotheses testing results were as follows:

1. There was no difference in measurement needs between elementary school teachers and secondary school teachers in Bangkok.
2. Certificate teachers had more measurement needs (indicated by lower score from the test) than degree teachers.
3. There was no difference in measurement needs between teachers who had more teaching experience and those who had less teaching experience.

4. There were some differences in measurement needs among the four subject matter areas, Planning a Classroom Test seemed to be the most needed (lowest mean).
5. There was an interaction between level of school and level of teacher education. There was no difference between the mean of certificate secondary school teachers and the mean of certificate elementary school teachers but degree elementary school teachers had more measurement needs than degree secondary school teachers.
6. There was no interaction between level of school and teaching experience.
7. There was no interaction between level of teacher education and teaching experience.
8. There was no three-way interaction among level of school, level of teacher education, and teaching experience.
9. There was no interaction between subject matter and level of school.
10. There was an ordinal interaction between subject matter and level of teacher education. The certificate teachers had more measurement needs in all subscales than degree teachers. The mean difference on Item Analysis subscale, however, was not a significant difference.
11. There was no interaction between subject matter and teaching experience.
12. There was no three-way interaction between subject matter, level of school, and level of teacher education.

13. There was no three-way interaction between subject matter, level of school, and teaching experience.
14. There was no three-way interaction between subject matter, level of teacher education, and teaching experience.
15. There was no four-way interaction between subject matter, level of school, level of teaching education, and teaching experience.

Since the subject matter by teacher education interaction was significant, interpretations of profile were made separately for certificate teachers and for degree teachers. Both groups of teachers got their lowest scores on Planning a Classroom Test subscale. The degree teachers got their highest scores on Test Score Statistics and Marking System subscale, but certificate teachers got their highest scores on the Item Analysis subscale. Degree teachers got higher scores than certificate teachers in all subscales. An interaction was found on the level of school by teacher education interaction. The total score mean indicated that degree teachers in secondary schools got higher scores than degree teachers in elementary schools, but the mean of certificate teachers in secondary schools was slightly lower than the mean of certificate teachers in elementary schools. There was no significant difference between certificate elementary school teachers and certificate secondary school teachers, but a significant difference between degree elementary school teachers and degree secondary school teachers was found at $\alpha = .05$ level.

Analyses were also done to compare measurement needs (indicated by lower scores from the test) among various groups of teachers (defined

by the following independent variables: taking a college measurement course, attending the training program in measurement, favoring a national testing program, and region of school), to observe the relationship between perceived needs and measurement needs, and to compare the mean differences between sample means and criterion scores.

Among teachers who took and did not take a college measurement course, the former group had less measurement needs (indicated by higher scores from the test) than those who did not take a course. There were no significant differences on measurement needs among teachers who attended the training program and those who did not attend the program, among teachers who favored and did not favor the national testing program, or among teachers who taught in five different regions. A relationship between perceived needs (indicated by the feeling of lacking desirable knowledge in measurement) and measurement needs was not found.

There were significant differences between means of the sample and criterion scores both on total mean and subscale means. The results showed that the teachers in Bangkok had measurement needs in all four subject matter areas.

CHAPTER V

SUMMARY AND CONCLUSIONS

Summary

This study was aimed at providing data concerning the quality of the teacher as an evaluator to the administrators and educators in Thailand. It was the purpose of the study to find out which groups of teachers actually need in-service training and in which areas of measurement the need is the greatest. This study also yields some follow-up information on the effects of the previous in-service programs in measurement and the effects of a measurement course offered by the teacher-training institutions.

The population of interest of this study was the public elementary and secondary school teachers, who are under the Ministry of Education in Bangkok. The instrument used in the study was a questionnaire concerning the teachers' opinions on national testing and their perceived needs in measurement and a true-false test measuring basic knowledge on educational measurement. The items were selected from the items used in a basic measurement course taught at Michigan State University. The items in the pilot test covered basic knowledge in measurement and evaluation corresponding to the four subject matters that teachers should know. The four areas are: planning a classroom test, item writing, item analysis, and test score statistics and marking

system. The pilot test was composed of 20 items in each subject matter area, with a total of 80 items. The instrument was translated into Thai and it was piloted out on forty Thai elementary school teachers and forty Thai secondary school teachers for a total of eighty teachers. The reliability of the pilot test was .40, mean item difficulty (percent of individuals giving an incorrect answer) was .53, and mean item discrimination was .15. The items for each subscale were then selected separately for the final test. Within each of the four subject matters, the item discrimination indices and item difficulties were considered in deleting items. The final test contained fifty-two items, thirteen items for each subscale.

The final instrument was composed of two parts:

1. A questionnaire concerning the teachers' opinions on national testing and their perceived needs in measurement. This part contained eight items.
2. Test items measuring basic knowledge in measurement and evaluation. The second part was composed of 52 true-false items of which 20 statements were true and 32 statements were false. The mean item difficulty for Planning a Classroom Test subscale was .50 (varied between .19 to .76), for the Item Writing subscale it was .49 (varied between .16 to .81), for the Item Analysis subscale it was .49 (varied between .09 to .83), and it was .55 (varied between .16 to .80) for the Test Score Statistics and Marking System subscale.

The final instrument was sent to 540 Thai teachers who were randomly selected from twelve strata. The stratification was based on the three variables: level of school - elementary school or secondary school; level of teacher education - teaching certificate holders or bachelor's degree holders; and teaching experience - less than or equal to three years, between four to ten years, or more than ten years. Because of the personal contacts and some help from the school principals, 69% of the responses (374 responses) were returned. Since the number of the returned responses in each group varied between 30 to 33, fourteen responses (3.7% of responses) were randomly thrown out in order to get thirty subjects in each group for the total of 360 subjects. Because of the homogeneity of the population and the fact that only 3.7% of the responses were randomly thrown out, it is believed that any distortion of data is small.

The design of this study was a 2 x 2 x 3 factorial design with four repeated measures. The design was crossed and balanced with 30 observations per cell. The multivariate repeated measures analysis was employed to test the research hypotheses.

Since the interaction between subject matter and teacher education was significant, profile interpretations were made separately for certificate teachers and for degree teachers. Both groups of teachers received their lowest scores on the Planning a Classroom Test subscale. The degree teachers received their highest scores on the Test Score Statistics and Marking System subscale, but certificate teachers received their highest scores on the Item Analysis subscale. Degree teachers got higher scores than certificate teachers in every subscale.

An interaction between the level of school by teacher education was also significant. The mean of the total score indicated that degree teachers in secondary school received a higher mean score than degree teachers in elementary school, but the mean of certificate teachers in secondary school was slightly lower than the mean of certificate teachers in elementary school. There was no significant difference between certificate elementary school teachers and certificate secondary school teachers, but a significant difference between degree teachers in elementary school and degree teachers in secondary school was found at $\alpha = .05$ level.

The F ratio for testing two hypotheses concerning the main effects were significant. There were teacher education main effects and the subject matter main effects. The data from Table 4.1 showed that the certificate teachers had more measurement needs than the degree teachers. It also indicated that measurement needs on Planning a Classroom Test was the highest need, and Item Analysis was the lowest need. However, the general profile could not be made applicable to all groups of teachers because the two-way interaction was significant.

Further comparisons were done to find if there were any differences among various groups of teachers (defined by the following independent variables: taking a college measurement course, attending the training program in measurement, favoring a national testing program, and region of school). Comparisons were also done to observe the relationship between perceived needs and measurement needs, and to compare the mean differences between sample means and criterion scores.

The F ratio from Table 4.5 indicated that the teachers who took a college measurement course had less total measurement needs (indicated by higher total scores from the test) than those who did not take a course. The same observations were true for the measurement needs in the Item Writing subscale, in the Item Analysis subscale, and in the Test Score Statistics and Marking System subscale, but not for the Planning a Classroom Test subscale. Although the teachers who took a college measurement course received a higher score on Planning a Classroom Test subscale than those who did not take a course, the difference was not large enough to be significant. It was found that there was no significant difference on measurement needs among teachers who attended the training program in measurement and those who did not attend the program. No significant difference was found among teachers who favored and did not favor the national testing program. It was also observed that there was not a significant difference between the teachers who taught in five different regions. The relationship between perceived needs (indicated by the feeling of lacking desirable knowledge in measurement) and measurement needs (indicated by lower test score) was not found.

There were significant differences between the means of the entire sample and the criterion scores on both the total mean and subscale means. The analyses indicated that the teachers in Bangkok had measurement needs in all four subject matter areas.

Conclusions and Implications

A cross-tabulation between took/did not take a college measurement course and attended/did not attend the training program in measurement and evaluation showed that eighty-five of 107 teachers who attended the training program took a college measurement course, 22 teachers did not take a course. It was observed that only 19% (69 teachers) of the total sample neither took a college measurement course nor attended the training program in measurement and evaluation.

There was a significant difference in measurement needs in favor of the teachers who took a college measurement course as compared to those who did not take a course, and the results of the study also indicated that the former group had less measurement needs than the latter group in Item Writing, Item Analysis, and Test Score Statistics and Marking System. It was found that there was no significant difference in test scores between those who took a measurement course and those who did not take a course on Planning a Classroom Test subscale, suggesting that this area of measurement might not have been included in the content of the college measurement courses. The results of the study seem to indicate that it would be appropriate to emphasize or include Planning a Classroom Test area in future college measurement courses.

Because of the interaction effect between teacher education and subject matter, the measurement needs for each group of teachers were different. Therefore, the subject matter should be arranged according to the needs of a majority of teachers in each training session. The results of this study indicated that the teachers who hold a teaching certificate had more measurement needs than those who hold at least a

bachelor's degree in every subject matter area. The area with the highest measurement needs for both certificate teachers and degree teachers was Planning a Classroom Test. The teachers who hold a teaching certificate had the lowest measurement needs in the Item Analysis area, but those who hold at least a bachelor's degree had the lowest measurement needs in Test Score Statistics and Marking System area. These seem to indicate that the degree teachers had more mathematics background than the certificate teachers, and because of the nature of the subject matter of measurement, with some mathematics involved, the holders of teaching certificates may turn down the invitation to join the training program or to take the measurement course. It is strongly recommended that an introductory course in educational measurement should be a requirement in the curriculum of the two-year and four-year teacher training program.

Although the elementary school teachers who hold a teaching certificate received a lower mean score than those who hold at least a bachelor's degree in every subscale, the amount of the differences was not large enough to be significant. For secondary school teachers, however, significant differences between those who hold at least a bachelor's degree and those who hold a teaching certificate were found in every subscale, except on the Item Analysis subscale. These results may suggest that the future in-service training program in measurement should be arranged for the elementary school teachers separately from secondary school teachers, and within secondary school teachers, the training program should be arranged for the degree teachers separately from the certificate teachers. The study also suggests that the

elementary school teachers who hold at least a bachelor's degree had more measurement needs than the secondary school teachers who hold the same level of education. This may be the result of a lack of interest or because of the heavy teaching loads. Most of the Thai elementary school teachers taught all subjects and for thirty hours a week. They might have little time to study or pay attention to other professional activities.

In testing the difference between teachers who attended and did not attend the in-service training program in measurement, no significant difference was found between these two groups. The result was supported by a previous study conducted by Sor-Wasna Pravalpruk (1974). She found that there was no significant difference between the teachers in Khon Kaen, Thailand who had attended and had not attended the in-service program in measurement. This result was probably caused by two factors. First, some of those who did not attend the training program had taken a college measurement course. Another factor was that the teachers had attended in-service programs of limited duration. In the past, most of the in-service programs in measurement and evaluation in Thailand were five-day workshops. The material covered purposes of measurement, curriculum analysis as a blueprint for test construction, types of test items, item analysis, scores and norms, and reporting the test results. The morning sessions were lectured by specialists in measurement, the afternoon sessions were practicums. It is recommended that the period of the future in-service training program should be longer than five days so that the teachers can have enough time to practice and learn the material.

The significant differences between the means of the entire sample and criterion scores indicated that the abilities of the Thai teachers in measurement and evaluation were lower than standard. The teachers in Bangkok had measurement needs in all four subject matter areas. Well-organized in-service programs should be offered to those teachers to increase the skills of development of teachers or to prepare teachers for new experiences in measurement and evaluation. A short course in measurement should be offered for the short-term effect. For the long-term effect, however, the teacher-training institutions should have full responsibility for improving the competence of the teachers in measurement. The teacher-training curriculum should be re-considered, and the contents of a measurement course should be revised. If continuing professional growth is to be taken seriously, administrators and teachers must pool their knowledge and resources and seek to make the in-service program and a college measurement course more responsive to the needs and interests of practicing classroom teachers. It would be expected that these programs might be useful to help school personnel become more familiar with test construction and evaluation.

Recommendations for Further Study

The previous in-service training program seemed to yield little benefit to the teachers in Bangkok. Perhaps this was because the design of the training program did not provide the functions necessary to meet the needs of the participants. Since a well-designed survey could serve as a learning experience for participants, a survey study should be done to provide information for planning any future training program

in measurement to achieve the expected outcomes. The questionnaire should be sent to the representatives of teachers and organizations in Bangkok to discover the current attitudes about the training in Bangkok, about needs of the participants, and to identify existing resources. The questions in the questionnaire might be divided into six categories as follows:

1. Attitudes toward popular participation in program design and implementation. How extensively should administrators be involved in the design of programs to upgrade their skills? Who would they select to design a program?
2. Previous experience with training programs and attitudes toward the training program. Which other training programs have they attended? Was the training worthwhile? What type of training do they think is most useful?
3. Content of the training program. Measurement needs might be discovered by the test items.
4. Format of the training program. How long should the training program last and where should it be located? Should the training program be offered during a single session?
5. Resources. Who should conduct the training program? What skills should a trainer have?
6. Techniques and materials. What types of learning situations do they prefer (informal discussion, lecture, workshop, etc.)? What types of support materials would be most useful to them?

Another study might be done to investigate the benefit from the training service. Any gain of knowledge after the training should be

studied to compare the gain made by the teachers who hold a teaching certificate with the gain made by the teachers who hold at least a bachelor's degree. Pretest and post-test procedures should be used to investigate if there is any measurement growth after the teachers have participated in the program.

A follow-up study on the quality of the teacher-made tests should also be done. If there is no improvement in the quality of the teacher-made tests, it might be wasted effort to offer the in-service training program in measurement to the teachers.

The quality of an introductory course in educational measurement should also be investigated. The study might be done by mailing a questionnaire to all teacher-training institutions in Thailand to examine whether the contents of a measurement course correspond to the needs and interests of the classroom teachers.

A national Test Bureau should be established, to be a center of testing services and to carry on a national testing program and other educational testing programs. Standardized (both achievement and aptitude) tests should be developed for the purpose of guidance, selection, and diagnosis of student learning, and should be available to all teachers and school personnel. National norms and local norms for the standardized tests should also be constructed to allow for further interpretation of test results.

APPENDIX A

THE INSTRUMENT
(FIRST EDITION)

APPENDIX A

THE INSTRUMENT

(FIRST EDITION)

Directions:

This test consists of 80 statements about basic knowledge in measurement and evaluation. You are to decide whether each statement is true or false. Please write the letter "T" in front of the true statements and "F" in front of the false statements.

You do not need to identify yourself, but please answer each of the test items as accurately and as honestly as you can. There is no time limit in answering the test items.

Part I: Planning a Classroom Test

1. Useful measurements are necessarily objective.

2. A teacher's skill in constructing tests for a subject depends more upon his general skill in test construction than it does on the quality of his knowledge of that subject.

3. The aspects of achievement that multiple-choice tests can measure are more limited than is the case for short-answer tests.

4. The use of a variety of item types in an examination is likely to improve the validity of the examination.

5. The most valid classroom tests of achievement tend to be those that most students have time to finish.

6. Sampling errors tend to be less serious in essay tests than in objective tests.

7. The choice between essay or objective test forms should be made primarily on the basis of class size.

8. Since students show a wide range of individual differences, the ideal measurement situation would be achieved if each student could take a different test that was specially designed to test him.

9. True-false test items are easier to write but less efficient than multiple-choice test items.

10. To obtain objective measurement of achievement, it is necessary to use objective test items.

11. If 240 items are available for measuring achievement in a course, a more reliable composite measure of achievement is likely to be obtained if these items are administered at different times as three separate 80-item tests than if they are all administered at the same time as a single test.

12. The number of items to be included in a test should be determined primarily by the amount of material the test must cover.

13. A one-hour objective test ordinarily provides a more extensive sample of a student's achievements than a one-hour essay test.

14. Frequent testing is more beneficial in the lower grades than it is in high school or college.

15. One should choose among essay, true-false, multiple-choice and other item forms depending on the particular mental ability that is to be tested.

16. Good achievement tests include approximately equal numbers of very easy, easy, average, difficult, and very difficult items.

17. Experts agree that cheating can be eliminated by the use of open-book.

18. Either too little or too much testing can lead to unreliable measurements of achievement.

19. Individual differences are more clearly apparent when all students take the same test than when each takes a test specially designed to test him.

20. A test composed entirely of items of moderate difficulty (neither very easy nor very hard) can nevertheless discriminate well among the very best students and among the very poorest students.

Part II: Item Writing

21. To be appropriate for inclusion in an achievement test, an item should deal with an idea emphasized in instruction.

22. The item writer should seek to prevent a student from getting the correct answer by a process of eliminating incorrect answers.

23. If textbook wording is followed closely in phrasing multiple-choice test items, students may be able to respond correctly without understanding.

24. The distractors in a multiple-choice item should be plausibly attractive but definitely incorrect.

25. The response "None of the above" makes a good fourth or fifth response to almost any multiple-choice test item.

26. In order to discriminate properly, a multiple-choice test item must provide at least four alternative responses (possible answers).

27. Making some questions optional tends to improve the reliability of essay test scores.

28. Almost any good true test item can be converted into an equally good false item simply by inserting the word "not" in it.

29. Most of the sentences in a well written textbook could be used as the true statements in a true-false test.

30. True statements that do not provide good answers to the stem question often make good distractors.

31. If a question can not be given an absolutely correct answer, it should not be included in an achievement test.

32. It is better for an item writer to review the items he has written after several days have passed, than ask someone else to review them.

33. Good multiple-choice items can be written using only the correct answer and one incorrect alternative.

34. If a response is stated more carefully, and at greater length than the other responses in a multiple-choice test item, the chances are that it is the correct response.

35. Multiple-choice items which ask the student to pick one incorrect answer from among several correct answers tend to be highly discriminating.

36. Multiple-choice items whose stems are stated negatively tend to be more discriminating than those whose stems are stated positively.

37. The item writer should aim to produce items that will be answered correctly by most students of high achievement, and missed by most students of low achievement.

38. Multiple-choice test items that call for only a "best" answer, instead of a perfectly correct answer, tend to be less discriminating and more ambiguous.

39. The responses "All of the above," or "None of the above," are recommended for use in almost all multiple-choice test items.

40. Multiple-choice items can be converted to equally effective true-false items in almost all cases.

Part III: Item Analysis

41. A "medium difficulty" true-false item is answered correctly less often than a "medium difficulty" multiple-choice item.

42. If nine of ten students who score high on a test answer a particular item correctly, while two of ten who score low on the test answer it correctly, the index of discrimination is .70.

43. Item analysis is more useful to a teacher who re-uses items than to one who does not.

44. If extreme groups of 33% instead of 27% are used for item analysis the groups will be more alike in average ability.

45. A wide distribution of item difficulty values in a test is likely to lead to a wide distribution of pupil scores on the test.

46. Most item analyses are based on external criterion measures of what the test is designed to measure.

47. If 12 of 20 students answer a question correctly, all 12 of them should be expected to answer another, easier question correctly.

48. Item analysis data can help the item writer identify and correct sources of weakness in a multiple-choice test item.

49. It is reasonable to regard most objective test items whose indices of discrimination are above .30 as weak and in need of improvement.

50. If six of ten students who score high on a test answer a particular item correctly, while four of ten who score low on the test answer the same item correctly, the index of difficulty of the item is .50.

51. The main reason for using upper and lower groups each including 27% rather than 50% of the total group tested is to reduce the labor of counting responses.

52. If three-fourths of the examinees who take a test answer an item correctly, its index of discrimination is .75.

53. If an item is extremely easy it is likely to be low in discrimination.

54. To determine the index of discrimination of a test item one must first determine its index of difficulty.

55. Ordinarily test papers must be scored before the items can be analyzed.

56. In general, the more difficult an item in a classroom test the higher its power of discrimination is likely to be.

57. The primary goal of item selection, on the basis of indices of discrimination, is to increase test reliability.

58. It is better to select the criterion groups used in item analysis at random than on the basis of total test score.

59. If the scores on Test A are much more variable than the scores on Test B, the difficulty values for the items in Test A are also likely to be more variable than those for the items in Test B.

60. Good classroom test items should have indices of discrimination of .50 or more.

Part IV: Test Score Statistics and Marking Systems

61. In a frequency distribution of scores for which the mean is 78 and the median is 65, there must be more extremely high scores than extremely low scores.

62. If two sets of scores have different variances, they must have different standard deviations.

63. More than half of the scores in a typical distribution are located more than one standard deviation away from the mean.

64. In a set of test scores there are three scores in the fifties: 51, 53, and 59. The percentile ranks of scores 51 and 53 will be more nearly the same than the percentile ranks of scores 53 and 59.

65. If a student's raw score on Test A is larger than his raw score on Test B, his percentile rank on Test A should be larger also.

66. When scores on a test are converted to stanines, some pupils are likely to get stanine scores of -3.5.

67. It is possible to get a correlation coefficient of +1.20.

68. For a group of nine year-olds, the correlation between age in years and I.Q.'s will be precisely zero.

69. Differences from instructor to instructor in marking are inevitable and educationally desirable.

70. It is better for a marking system to report absolute than relative achievement.

71. Percentage marks were intended to report the proportion learned of that which might have been learned.

72. Increasing the number of categories of marks tends to increase the reliability of the marks.

73. If a set of eight scores includes two eights -- two sevens -- two fives and two fours, the median value is six.

74. The distribution of the scores 5, 4, 3, 2, 1 is approximately normal.

75. If in a distribution of 100 scores there are four scores of 28 and 30 scores lower than 28, the percentile rank of 28 is 32.

76. When students are grouped by ability levels the policy of giving a higher proportion of A's in the more able group is justifiable.

77. By using fewer, broader categories in marking a teacher can reduce the proportion of incorrect marks he issues without seriously reducing the amount of useful information he reports.

78. Stanine marks are likely to be more reliable than five-letter marks.

79. No instructor is entitled to criticize the distribution of marks in another instructor's course.

80. The distribution of marks in all classes should be approximately the same regardless of differences in the general levels of ability of the students in the different classes.

APPENDIX B

THE INSTRUMENT
(FINAL EDITION)

APPENDIX B

THE INSTRUMENT

(FINAL EDITION)

Directions:

You do not need to identify yourself. Please answer each of the questions on this page and the following pages as accurately and as honestly as you can. There is no time limit in answering this questionnaire.

PART I: General Information.

Please check the appropriate categories.

1. Level of school you teach:

   /____/ Elementary school

   /____/ Secondary school

2. Level of your education:

   /____/ Certificate or lower

   /____/ Bachelor degree or higher

3. Teaching experience:

   /____/ 0 - 3 years

   /____/ 4 - 10 years

   /____/ More than 10 years

4. Did you take any measurement course when you studied in college?

   /____/ Yes /____/ No

5. Did you attend in-service training programs in measurement and evaluation?

/____/ Yes /____/ No

If the answer is "yes," go to question 5.1

If the answer is "no," go to question 5.2

5.1. Was the training program worthwhile?

/____/ Yes /____/ No

5.2. If the Ministry of Education offers the training program in measurement and evaluation, will you participate in that program?

/____/ Yes /____/ No

6. Are you in favor of national testing?

/____/ Yes /____/ No

7. What area of measurement do you know most? (check only one)

/____/ Planning a classroom test

/____/ Item writing

/____/ Item analysis

/____/ Test score statistics and marking systems

8. What area of measurement do you know least? (check only one)

/____/ Planning a classroom test

/____/ Item writing

/____/ Item analysis

/____/ Test score statistics and marking systems

PART II: Basic Knowledge in
Measurement and Evaluation

The following are statements about basic knowledge in measurement and evaluation. You are to decide whether each statement is true or false. Please write the letter "T" in front of the true statements and "F" in front of the false statements.

Please try to answer all of the 52 statements.

1. It is necessary to use different test forms to test different abilities.

2. The aspects of achievement that multiple-choice tests can measure are more limited than is the case for short-answer tests.

3. The use of a variety of item types in an examination is likely to improve the validity of the examination.

4. The most valid classroom tests of achievement tend to be those that most students have time to finish.

5. Sampling errors tend to be less serious in essay tests than in objective tests.

6. Since students show a wide range of individual differences, the ideal measurement situation would be achieved if each student could take a different test that was specially designed to test him.

7. To obtain objective measurement of achievement, it is necessary to use objective test items.

8. If 240 items are available for measuring achievement in a course, a more reliable composite measure of achievement is likely to be obtained if these items are administered at different times as three separate 80-item tests than if they are all administered at the same time as a single test.

9. A one-hour objective test ordinarily provides a more extensive sample of a student's achievements than a one-hour essay test.

10. Frequent testing is more beneficial in the lower grades than it is in high school or college.

11. One should choose among essay, true-false, multiple-choice and other item forms depending on the particular mental ability that is to be tested.

12. Good achievement tests include approximately equal numbers of very easy, easy, average, difficult, and very difficult items.

13. A test composed entirely of items of moderate difficulty (neither very easy nor very hard) can nevertheless discriminate well among the very best students, and among the very poorest students.

14. To be appropriate for inclusion in an achievement test an item should deal with an idea emphasized in instruction.

15. If textbook wording is followed closely in phrasing multiple-choice test items, students may be able to respond correctly without understanding.

16. The distractors in a multiple-choice item should be plausibly attractive but definitely incorrect.

17. The response "None of the above" makes a good fourth or fifth response to almost any multiple-choice test item.

18. Making some questions optional tends to improve the reliability of essay test scores.

19. Most of the sentences in a well written textbook could be used as the true statements in a true-false test.

20. True statements that do not provide good answers to the stem question often make good distractors.

21. If a question can not be given an absolutely correct answer, it should not be included in an achievement test.

22. It is better for an item writer to review the items he has written after several days have passed, than ask someone else to review them.

23. Multiple-choice items which ask the student to pick one incorrect answer from among several correct answers tend to be highly discriminating.

24. Multiple-choice items whose stems are stated negatively tend to be more discriminating than those whose stems are stated positively.

25. The item writer should aim to produce items that will be answered correctly by most students of high achievement, and missed by most students of low achievement.

26. The responses "All of the above," or "None of the above," are recommended for use in almost all multiple-choice test items.

27. A "medium difficulty" true-false item is answered correctly more often than a "medium difficulty" multiple-choice item.

28. If extreme groups of 33% instead of 27% are used for item analysis the groups will be more alike in average ability.

29. A wide distribution of item difficulty values in a test is likely to lead to a wide distribution of pupil scores on the test.

30. If 12 of 20 students answer a question correctly, all 12 of them should be expected to answer another, easier question correctly.

31. Item analysis data can help the item writer identify and correct sources of weakness in a multiple-choice test item.

32. If six of ten students who score high on a test answer a particular item correctly, while four of ten who score low on the test answer the same item correctly, the index of difficulty of the item is .50.

33. The main reason for using upper and lower groups each including 27% rather than 50% of the total group tested should be to reduce the labor counting responses.

34. If three-fourths of the examinees who take a test answer an item correctly, its index of discrimination is .75.

35. If an item is extremely easy it is likely to be low in discrimination.

36. To determine the index of discrimination of a test item one must first determine its index of difficulty.

37. The primary goal of item selection, on the basis of indices of discrimination, is to increase test reliability.

38. If the scores on Test A are much more variable than the scores on Test B, the difficulty values for the items in Test A are also likely to be more variable than those for the items in Test B.

39. Good classroom test items should have indices of discrimination of .50 or more.

40. In a frequency distribution of scores for which the mean is 78 and the median is 65, there must be more extremely high scores than extremely low scores.

41. If two sets of scores have different variances, they must have different standard deviations.

42. More than half of the scores in a typical distribution are located more than one standard deviation away from the mean.

43. If a student's raw score on Test A is larger than his raw score on Test B, his percentile rank on Test A should be larger also.

44. When scores on a test are converted to stanines, some pupils are likely to get stanine scores of -3.5.

45. Differences from instructor to instructor in marking are inevitable and educationally desirable.

46. It is better for a marking system to report absolute than relative achievement.

47. Percentage marks were intended to report the proportion learned of that which might have been learned.

48. If two sets of scores have different means they must have different variances.

49. When students are grouped by ability levels the policy of giving a higher proportion of A's in the more able group is justifiable.

50. Stanine marks are likely to be more reliable than five-letter marks.

51. No instructor is entitled to criticize the distribution of marks in another instructor's course.

52. The distribution of marks in all classes should be approximately the same regardless of differences in the general levels of ability of the students in the different classes.

แบบสอบถามความรู้ทั่วไปเกี่ยวกับการวัดผลและการประเมินผลการศึกษา

คำชี้แจง

แบบสอบถามฉบับนี้แบ่งออกเป็น ๒ ตอน คือ ตอนที่ ๑ เป็นรายละเอียดเกี่ยวกับผู้ตอบมี ๘ ข้อ และตอนที่ ๒ เป็นความคิดเห็นด้านความรู้ทั่วไปเกี่ยวกับการวัดผลและการประเมินผลการศึกษา มี ๔๒ ข้อความ

ขอความกรุณาท่านอาจารย์ตอบคำถามทุกข้อด้วย โดยไม่ต้องเขียนชื่อและนามสกุลลงในแบบสอบถามและไม่มีกำหนดเวลาสำหรับการตอบแบบสอบถามฉบับนี้

"ขอขอบพระคุณเป็นอย่างสูงในการให้ความร่วมมืออย่างดียิ่ง"

ตอนที่ ๑ รายละเอียดเกี่ยวกับผู้ตอบ โปรดกรุณาเขียนเครื่องหมาย " ✓ " ลงใน ☐ หน้าข้อความที่ตรงกับสภาพความเป็นจริงของท่าน

---

๑. ระดับชั้นที่ท่านสอน

☐ ระดับประถมศึกษา

☐ ระดับมัธยมศึกษา

๒. วุฒิของท่าน

☐ อนุปริญญาหรือต่ำกว่า

☐ ปริญญาตรีหรือสูงกว่า

๓. เวลาที่รับราชการ

☐ ๓ ปีและน้อยกว่า

☐ ๔ - ๑๐ ปี

☐ ๑๑ ปีและมากกว่า

๔. ท่านเคยเรียนวิชาการวัดผลการศึกษาจากวิทยาลัยหรือมหาวิทยาลัยหรือไม่?

☐ เคย

☐ ไม่เคย

๕. ท่านเคยเข้ารับการอบรมวิชาการวัดผลการศึกษาหรือไม่?

☐ เคย

☐ ไม่เคย

ถ้าตอบว่า "เคย" ให้ไปตอบข้อ ๕.๑ ถ้าตอบว่า "ไม่เคย" ให้ไปตอบข้อ ๕.๒

๕.๑ ท่านได้ประโยชน์จากการเข้ารับการอบรมวิชาการวัดผลการศึกษาหรือไม่?

☐ ได้

☐ ไม่ได้

๕.๒ ถ้ากระทรวงศึกษาธิการจะเปิดการอบรมการวัดผลการศึกษาอีก ท่านจะสมัครเข้ารับการอบรมหรือไม่?

☐ สมัคร

☐ ไม่สมัคร

๖. ท่านชอบการสอบโดยใช้ข้อสอบรวมหรือไม่? ☐ ชอบ ☐ ไม่ชอบ ๗. ท่านคิดว่าท่านมีความรู้เกี่ยวกับการวัดผลในเรื่องใดมากที่สุด (เลือกตอบเพียงข้อเดียว) ☐ การวางแผนการสร้างข้อสอบ ☐ การเขียนข้อสอบ ☐ การวิเคราะห์ข้อสอบ ☐ สถิติเกี่ยวกับคะแนนและการตัดเกรด	๘. ท่านคิดว่าท่านมีความรู้เกี่ยวกับการวัดผลในเรื่องใดน้อยที่สุด (เลือกตอบเพียงข้อเดียว) ☐ การวางแผนการสร้างข้อสอบ ☐ การเขียนข้อสอบ ☐ การวิเคราะห์ข้อสอบ ☐ สถิติเกี่ยวกับคะแนนและการตัดเกรด

ตอนที่ ๒ ความรู้ทั่วไปเกี่ยวกับการวัดผลและการประเมินผลการศึกษา ขอความกรุณาท่านอาจารย์อ่านข้อความแต่ละข้อ และตัดสินใจว่าข้อความนั้นเป็นจริงหรือเป็นเท็จ ถ้าข้อความนั้นเป็นจริง ให้เขียนเครื่องหมาย " ✓ " ลงในวงเล็บหน้าข้อความนั้น และถ้าข้อความนั้นเป็นเท็จ ให้เขียนเครื่องหมาย " X " ลงในวงเล็บหน้าข้อความนั้น

ตัวอย่างการตอบ

(X) ๐. ครูไม่จำเป็นต้องประกาศวันสอบให้นักเรียนรู้ตัวล่วงหน้า

(✓) ๐๐. การวัดผลที่ดีจะต้องมีความเป็นปรนัย

(...) ๑. การวัดความสามารถที่แตกต่างกันจำเป็นต้องใช้แบบของคำถามที่แตกต่างกัน (เช่นคำถามแบบอัตนัย คำถามแบบเลือกตอบ คำถามแบบถูก-ผิด ฯลฯ)

(...) ๒. ข้อสอบแบบเลือกตอบวัดเนื้อหาวิชาได้แคบกว่าข้อสอบแบบเติมคำหรือเติมข้อความสั้น ๆ (short answer tests)

(...) ๓. การใช้คำถามหลาย ๆ แบบ (เช่นแบบเลือกตอบ แบบถูกผิด แบบจับคู่) ในการสอบครั้งหนึ่ง เป็นการปรับปรุงความเที่ยงตรง (validity) ของการทดสอบ

(...) ๔. ข้อสอบวัดผลสัมฤทธิ์ที่มีความเที่ยงตรงสูง มักจะเป็นข้อสอบที่ให้เวลามากพอจนกระทั่งนักเรียนส่วนใหญ่มีเวลาทำข้อสอบจนเสร็จ

(...) ๕. ข้อสอบอัตนัย (essay tests) วัดเนื้อหาได้ครอบคลุมกว่าข้อสอบปรนัย (objective tests)

(...) ๖. เนื่องจากนักเรียนแต่ละคนมีความแตกต่างกันมาก การวัดผลจะมีประสิทธิภาพ ถ้านักเรียนได้สอบข้อสอบซึ่งออกเป็นพิเศษเฉพาะสำหรับเขา

(...) ๗. ครูควรจะใช้ข้อสอบปรนัย ถ้าต้องการให้เกิดความเป็นปรนัยในการวัดผล

(...) ๘. ถ้ามีข้อสอบสำหรับวัดผลสัมฤทธิ์ทางการเรียนวิชาหนึ่งอยู่ ๒๔๐ ข้อ การวัดผลจะมีความเชื่อมั่น (reliability) ได้มากขึ้น ถ้าครูแบ่งข้อสอบ ๒๔๐ ข้อ ออกเป็นข้อสอบ ๓ ฉบับ ฉบับละ ๘๐ ข้อ แล้วดำเนินการสอบ ๓ ครั้ง ในเวลาที่แตกต่างกัน แทนที่จะสอบครั้งเดียว ๒๔๐ ข้อ

(...) ๙. ข้อสอบปรนัยที่ให้เวลาทำ ๑ ชั่วโมง จะสามารถวัดผลสัมฤทธิ์ทางการเรียนของนักเรียนได้มากกว่าข้อสอบอัตนัยที่ให้เวลาทำ ๑ ชั่วโมงเช่นกัน

(...) ๑๐. การสอบบ่อย ๆ มีประโยชน์สำหรับนักเรียนชั้นประถมมากกว่านักเรียนชั้นมัธยมหรือนักศึกษาในมหาวิทยาลัย

(...) ๑๑. การที่จะตัดสินใจว่าควรจะออกข้อสอบแบบอัตนัยหรือแบบเลือกตอบ หรือแบบถูก-ผิด นั้นขึ้นอยู่กับว่าจะวัดสมรรถภาพสมองทางด้านใด

(...) ๑๒. ข้อสอบวัดผลสัมฤทธิ์ที่ดี ควรจะมีจำนวนข้อของข้อสอบที่ง่ายมาก ง่าย ปานกลาง ยากและยากมาก มีจำนวนพอ ๆ กัน

(...) ๑๓. ข้อสอบฉบับหนึ่งซึ่งประกอบไปด้วยคำถามที่มีความยากง่ายปานกลาง (ไม่ยากหรือไม่ง่ายเกินไป) ไม่สามารถนำมาใช้ในการจำแนกความสามารถระหว่างนักเรียนกลุ่มเก่งด้วยกันหรือระหว่างนักเรียนกลุ่มอ่อนด้วยกันได้

(...) ๑๔. ข้อสอบวัดผลสัมฤทธิ์ที่ดีควรจะวัดในสิ่งซึ่งเกี่ยวข้องกับเนื้อหาวิชาที่ครูสอน

(...) ๑๕. ถ้าคำหรือประโยคจากตำราเรียนถูกนำมาใช้ในข้อสอบแบบเลือกตอบคำต่อคำ หรือประโยคต่อประโยคแล้ว นักเรียนอาจจะสามารถตอบข้อสอบนั้นถูก โดยปราศจากความเข้าใจในเนื้อหานั้น ๆ

(...) ๑๖. ตัวลวง (คำตอบผิด) ในข้อสอบแบบเลือกตอบ ควรจะเป็นคำตอบที่ดึงดูดใจให้นักเรียนเลือกตอบ แต่โดยแท้จริงแล้วมันเป็นคำตอบที่ผิด

(...) ๑๗. คำตอบ "ไม่มีข้อใดถูก" เป็นคำตอบที่ครูควรนำมาใช้เป็นตัวเลือกที่ ๔ หรือที่ ๕ ในข้อสอบแบบเลือกตอบ

(...) ๑๘. การเขียนข้อสอบอัตนัยหลาย ๆ ข้อให้นักเรียนมีสิทธิเลือกทำเป็นการช่วยปรับปรุงความเชื่อมั่น (reliability) ของคะแนนให้สูงขึ้น

(...) ๑๙. ประโยคหรือข้อความส่วนใหญ่จากหนังสือที่ดี ๆ ที่ใช้เป็นตำราเรียน สามารถนำมาใช้เป็นข้อความที่เป็นจริง (true statement) ในข้อสอบแบบถูก-ผิดได้เป็นอย่างดี

(...) ๒๐. ประโยคหรือข้อความที่เป็นจริงซึ่งไม่ใช่คำตอบที่ถูกต้องของคำถามแบบเลือกตอบมักจะเป็นตัวลวงที่ดี

(...) ๒๑. ถ้าคำถามข้อหนึ่งไม่สามารถหาคำตอบที่ถูกต้องอย่างสมบูรณ์ได้ ผู้ที่ทำการออกข้อสอบควรจะตัดคำถามข้อนั้นทิ้งไป

(...) ๒๒. ผู้ที่ทำการเขียนข้อสอบ ควรจะทำการตรวจทานข้อสอบที่เขาเขียนขึ้นเองหลังจากที่เวลาผ่านไปแล้วหลาย ๆ วัน ดีกว่าที่จะขอร้องให้ผู้อื่นช่วยตรวจทานข้อสอบเหล่านั้น

(...) ๒๓. ข้อสอบแบบเลือกตอบซึ่งกำหนดให้นักเรียนเลือกคำตอบผิดจากคำตอบถูกหลาย ๆ คำตอบ มีแนวโน้มที่จะเป็นข้อสอบที่มีค่าอำนาจจำแนกสูง

(...) ๒๔. ข้อสอบแบบเลือกตอบซึ่งเขียนคำถามในรูปของประโยคปฏิเสธ มีแนวโน้มที่จะมีอำนาจจำแนกสูงกว่าข้อสอบแบบเลือกตอบ ซึ่งเขียนคำถามในรูปของประโยคบอกเล่าธรรมดา

(...) ๒๕. ครูควรจะมีจุดประสงค์ที่จะผลิตข้อสอบซึ่งนักเรียนกลุ่มเก่งตอบถูกเป็นส่วนใหญ่ และนักเรียนกลุ่มอ่อนตอบผิดเป็นส่วนใหญ่

(...) ๒๖. คำตอบประเภท "ถูกหมดทุกข้อ" หรือ "ไม่มีข้อใดถูก" ควรจะนำมาใช้เป็นตัวเลือกของข้อสอบแบบเลือกตอบ

(...) ๒๗. อาจจะมีคนตอบคำถามของข้อสอบแบบถูก-ผิดที่มีความยากง่ายปานกลาง ถูกเป็นจำนวนครั้งมากกว่าตอบคำถามของข้อสอบแบบเลือกตอบที่มีความยากง่ายปานกลางเหมือนกัน

(...) ๒๘. ถ้ากลุ่ม ๓๓% สูงสุด และกลุ่ม ๓๓% ต่ำสุดถูกนำมาใช้ในการวิเคราะห์ข้อสอบแทนที่จะใช้กลุ่ม ๒๗% สูงสุด และ ๒๗% ต่ำสุดแล้ว กลุ่ม ๓๓% สูงสุด-ต่ำสุด จะมีความสามารถเฉลี่ยใกล้เคียงกันมากกว่าความสามารถเฉลี่ยของกลุ่ม ๒๗% สูงสุด-ต่ำสุด

(...) ๒๙. ถ้าข้อสอบฉบับหนึ่งมีการกระจายของค่าความยากง่ายของข้อสอบกว้าง การกระจายของคะแนนของนักเรียนที่ได้จากการสอบข้อสอบฉบับนั้นจะกว้างตามไปด้วย

(...) ๓๐. ถ้านักเรียน ๑๒ คน จากจำนวนนักเรียนทั้งหมด ๒๐ คน ตอบคำถามข้อหนึ่งถูก นักเรียนทั้ง ๑๒ คน นั้นจะต้องตอบคำถามข้ออื่น ๆ ซึ่งง่ายกว่าคำถามข้อนั้นถูกด้วย

(...) ๓๑. ตัวเลขที่ได้จากการวิเคราะห์ข้อสอบสามารถช่วยให้ผู้ที่เขียนข้อสอบตรวจสอบและแก้ไขข้อบกพร่องของคำถามแบบเลือกตอบได้ตรงจุด

(...) ๓๒. ถ้า ๖ ใน ๑๐ คนของนักเรียนกลุ่มที่ได้คะแนนสูง (upper group) จากการสอบข้อสอบฉบับหนึ่งตอบคำถามข้อหนึ่งถูก ในขณะที่ ๔ ใน ๑๐ คนของนักเรียนกลุ่มที่ได้คะแนนต่ำ (lower group) ตอบคำถามข้อเดียวกันนั้นถูกแล้ว ข้อสอบข้อนั้นจะมีค่าความยากง่ายประมาณ .๕๐

(...) ๓๓. จุดมุ่งหมายใหญ่ของการใช้กลุ่ม ๒๗% สูง-ต่ำในการวิเคราะห์ข้อสอบแทนที่จะใช้กลุ่ม ๕๐% สูง-ต่ำก็เพื่อที่จะลดแรงงานในการนับจำนวนคำตอบและการคิดคำนวณ

(...) ๓๔. ถ้า ๓ ใน ๔ ของผู้เข้าสอบตอบคำถามข้อหนึ่งถูก ค่าอำนาจจำแนกของคำถามข้อนั้นจะเท่ากับ .๗๕

(...) ๓๕. ข้อสอบที่ง่าย ๆ มีแนวโน้มที่จะมีค่าอำนาจจำแนกต่ำ

(...) ๓๖. ครูจะต้องคำนวณหาค่าความยากง่ายของข้อสอบแต่ละข้อเสียก่อนจึงจะสามารถคำนวณค่าอำนาจจำแนกของข้อสอบเหล่านั้นได้

(...) ๓๗. จุดมุ่งหมายใหญ่ของการใช้ค่าอำนาจจำแนก (indicies of discrimination) เป็นเกณฑ์ในการคัดเลือกข้อสอบก็คือความต้องการที่จะเพิ่มค่าความเชื่อมั่นของข้อสอบให้สูงขึ้น

(...) ๓๘. ถ้าคะแนนที่ได้จากการสอบข้อสอบฉบับ ก. มีการกระจายมากกว่าคะแนนที่ได้จากการสอบข้อสอบฉบับ ข. แล้ว ค่าความยากง่ายของคำถามแต่ละข้อในข้อสอบฉบับ ก. จะมีการกระจายมากกว่าค่าความยากง่ายของคำถามแต่ละข้อในข้อสอบฉบับ ข.

(...) ๓๙. คำถามที่ดีควรจะมีค่าอำนาจจำแนกตั้งแต่ .๕๐ ขึ้นไป

(...) ๔๐. ถ้าคะแนนเฉลี่ย (mean) ของการสอบครั้งหนึ่งมีค่าเท่ากับ ๗๘. median มีค่าเท่ากับ ๖๔ แล้ว ในการสอบครั้งนี้จะมีจำนวนของนักเรียนที่สอบได้คะแนนสูงมากกว่าจำนวนของนักเรียนที่สอบได้คะแนนต่ำ

(...) ๔๑. ถ้าคะแนน ๒ ชุด มีความแปรปรวนมาตรฐาน (variance) ต่างกันแล้ว คะแนน ๒ ชุด นั้นจะต้องมีความเบี่ยงเบนมาตรฐาน (standard deviation) ต่างกันด้วย

(...) ๔๒. มากกว่าครึ่งหนึ่งของคะแนนในการกระจายปรกติ (normal distribution) จะตกอยู่ในสิ่นที่ตั้งแต่ +1ความเบี่ยงเบนมาตรฐานจากคะแนนเฉลี่ย (+1sd from mean) ขึ้นไป

(...) ๔๓. ถ้านักเรียนคนหนึ่งทำข้อสอบฉบับ ก. ได้คะแนนสูงกว่าข้อสอบฉบับ ข. แล้ว ตำแหน่งเปอร์เซ็นไทล์ที่ได้จากการสอบข้อสอบฉบับ ก. ของนักเรียนคนนั้นจะต้องสูงกว่าตำแหน่งเปอร์เซ็นไทล์ของเขาที่ได้จากการสอบข้อสอบฉบับ ข. ด้วย

(...) ๔๔. เมื่อคะแนนที่ได้จากการทดสอบฉบับหนึ่งถูกแปลงเป็นคะแนน stanine จะมีนักเรียนบางคนสอบได้คะแนน stanine ที่ - ๓.๔

(...) ๔๕. ความแตกต่างกันในหลักการของการให้คะแนน นอกจากจะเป็นสิ่งที่หลีกเลี่ยงไม่ได้แล้วยังเป็นสิ่งที่พึงปรารถนาทางการศึกษา

(...) ๔๖. ในการรายงานผลการเรียน ครูควรรายงานความสัมฤทธิ์ผลเฉพาะตัวของนักเรียน (เช่น วิชิตสอบพีชคณิตได้เกรด B) ดีกว่าที่จะรายงานความสัมฤทธิ์ผลของนักเรียนโดยการเปรียบเทียบกับความสามารถของนักเรียนคนอื่น ๆ (เช่น วิชิตสอบพีชคณิตได้เปอร์เซ็นไทล์ที่ ๗๐)

(...) ๔๗. การให้คะแนนเป็นเปอร์เซ็นเป็นความตั้งใจของครูที่จะรายงานสัดส่วนของการเรียนรู้ที่นักเรียนควรจะรู้

(...) ๔๘. ถ้าคะแนน ๒ ชุด มีค่าคะแนนเฉลี่ย (mean) ที่แตกต่างกัน คะแนน ๒ ชุดนั้นจะมีค่าความแปรปรวนมาตรฐาน (variance) ที่แตกต่างกันด้วย

(...) ๔๙. เมื่อนักเรียนถูกจัดกลุ่มตามระดับความสามารถ (เช่นนักเรียนห้อง ก. หมายถึงนักเรียนกลุ่มเก่ง นักเรียนห้อง ค. คือนักเรียนกลุ่มอ่อน) แล้วในการตัดเกรดวิชาใด ๆ ก็ตาม ถ้าครูตัดเกรดโดยพิจารณาคะแนนของนักเรียนเฉพาะแต่ละห้องโดยที่ไม่คำนึงถึงคะแนนของนักเรียนห้องอื่น ๆ แล้ว สัดส่วนของจำนวนนักเรียนที่ได้เกรด A ในกลุ่มเก่ง ควรจะได้รับการพิจารณาให้มากกว่า กลุ่มที่มีความสามารถรองลงมา

(...) ๕๐. การแบ่งเกรดเป็น ๔ เกรด จะทำให้เกรดเป็นที่เชื่อมั่นได้มากกว่าการแบ่งเกรดเป็น ๕ เกรด

(...) ๕๑. ไม่มีครูคนใดได้รับสิทธิ์ให้วิจารณ์การกระจายของเกรดในวิชาที่ครูคนอื่นทำการสอน

(...) ๕๒. การกระจายของเกรดที่ครูให้แก่นักเรียนแต่ละห้องควรจะเท่ากัน (เช่นนักเรียนห้อง ก. ได้ A ๕ คน นักเรียนห้อง ข. และห้อง ค. ก็ต้องได้ A ห้องละ ๕ คนด้วย) โดยไม่คำนึงว่าระดับความสามารถของนักเรียนแต่ละห้องจะเท่ากันหรือไม่

"ขอขอบพระคุณที่ท่านได้กรุณาตอบแบบสอบถามจนครบทุกข้อ"

BIBLIOGRAPHY

BIBLIOGRAPHY

Allen, M. E. "Status of measurement courses for undergraduates in teacher-training institutions." The 13th Yearbook of the National Council on Measurement in Education. 1956: 67-73.

Brimm, J. L. and D. J. Tollett. "How do teachers feel about in-service education." Educational Leaderships. 31 (March 1974): 521-525.

Conant, J. B. The Education of American Teachers. New York: McGraw-Hill Book Company, 1963.

Durost, W. N. "Problems in in-service training of teachers in the use of measurement and evaluation techniques." The 16th Yearbook of the National Council on Measurement in Education. 1959: 31-33.

Ebel, R. L. "Standardized achievement tests: Use and limitation." National Elementary School Principal. 40 (1961a): 29-32.

Ebel, R. L. "Improving the competence of teachers in educational measurement." Clearing House. 36 (October 1961b): 67-71.

Ebel, R. L. "Can teachers write good true-false test items?" Journal of Educational Measurement. 12 (Spring 1975): 31-35.

Ebel, R. L. Essentials of Educational Measurement. Prentice-Hall, Inc., 1979.

Erickson, R. C. and T. L. Wentling. Measuring Student Growth: Techniques and Procedures for Occupational Education. Allyn and Bacon, Inc., 1976.

Feldhusen, J. F. "Student perceptions of frequent quizzes and post-mortem discussions of tests." Journal of Educational Measurement. 1 (Jan. 1964): 51-54.

Finn, J. D. Finn's Multivariance - Univariate and Multivariate Analysis of Variance, Covariance, and Regression. Modified and adapted for use on the CDC-6500 at Michigan State University by Verda M. Scheifley and William H. Schmidt, 1973.

Fleming, Margaret. "Standardized tests revisited." School Counselor. 19 (November 1971): 71-72.

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

Lien, A. J. Measurement and Evaluation of Learning. W. M. C. Brown Company Publishers, 1971.

Lindvall, C. M. Measuring Pupil Achievement and Aptitude. Harcourt, Brace & World, Inc., 1967.

Marso, R. N. "Classroom testing procedures, test anxiety and achievement." Journal of Experimental Education. 38 (Spring 1970): 54-58.

Mayo, S. T. "Preservice preparation of teachers in educational measurement." Final report. Chicago: Loyola University, 1967.

Mehrens, W. A. "The technology of competency measurement." In R. B. Ingle, M. R. Carroll, & W. J. Gephart (Eds.) Assessment of Student Competence. Bloomington, Indiana: Phi Delta Kappa, 1979.

Mehrens, W. A. and I. J. Lehmann. Measurement and Evaluation in Education and Psychology. New York: Holt, Rinehart and Winston, Inc., 1978.

Miller, W. C. "What's wrong with in-service education? It's topless!" Educational Leadership. 35 (October 1977): 31-33.

Nie, N. H., C. Hull, J. Jenkins, K. Steinbrenner, and D. Bent. Statistical Package for the Social Sciences. McGraw-Hill Book Company, 1975.

Noll, V. H. "Requirements in educational measurement for prospective teachers." School and Society. 82 (September 1955): 88-90.

Noll, V. H. "Problems in the pre-service preparation of teachers in measurement." The 18th Yearbook of the National Council on Measurement in Education. 1961: 35-42.

Noll, V. H., D. P. Scannell, and R. C. Craig. Introduction to Educational Measurement. Houghton Mifflin Company, 1979.

Olejnik, S. F. "Standardized achievement programs viewed from the perspective of non-measurement specialist." Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, April, 1979.

Office of the National Education Commission. A Study of Primary Schooling in Thailand. Final report. Bangkok, Thailand, 1977.

Pravalpruk, S. Comparison among Teachers in Khon Kaen, Thailand, to Determine Their Testing Needs. Dissertation, Michigan State University, 1974.

Sack, W. M. Measurement Competencies of Educators Defined Through Task Analysis and Differentiated by Teaching Area, Grade Level, and Vocation. Dissertation, Loyola University of Chicago, 1979.

Stanley, J. C. Measurement in Today's Schools. Prentice-Hall, Inc., 1964.

Stanley, J. C. "ABC's of test construction." In T. M. Covin (Ed.) Classroom Test Construction. MSS Information Corporation, New York, 1974.

Stanley, J. C., and K. D. Hopkins. Educational and Psychological Measurement and Evaluation. Prentice-Hall, Inc., 1972.

Stetz, F. Report of a Market Survey to Sample Opinions of Sales Representatives and Users, Concerning the 1973 Stanford Achievement Test. Market Research Report No. 10. New York: The Psychological Corporation, 1977: 25.

Stevenson, M. "The role of the classroom teacher in school testing programs." The 16th Yearbook of the National Council on Measurement in Education. 1959: 43-46.

Thorndike, R. L., and E. Hagen. Measurement and Evaluation in Psychology and Education. John Wiley & Sons, Inc., 1969.

Yeh, J. P. "Test use in schools." Washington, D.C.: U. S. Department of Health, Education and Welfare, and National Institute of Education, 1978.

Zigami, P., L. Betz, and D. Jensen. "Teachers' preferences in and perceptions of in-service education." Educational Leadership. 34 (April 1977): 545-551.