This is to certify that the

dissertation entitled

Assessing the Competency of Teachers, Curriculum Specialists, and Prospective Teachers in Educational Measurement in Bahrain.

presented by

Rashid Hammad Aldosary

has been accepted towards fulfillment
of the requirements for

Ph.D. degree in Measurement, Evaluation, and Research Design.

Major professor

Date 6/21/93

PLACE IN RETURN BOX to remove this checkout from your record.
TO AVOID FINES return on or before date due.

DATE DUE	DATE DUE	DATE DUE

MSU Is An Affirmative Action/Equal Opportunity Institution

ASSESSING THE COMPETENCY OF TEACHERS, CURRICULUM SPECIALISTS, AND PROSPECTIVE TEACHERS IN EDUCATIONAL MEASUREMENT IN BAHRAIN

By

Rashid Hammad Aldosary

A DISSERTATION

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

DOCTOR OF PHILOSOPHY

Department of Counseling, Educational Psychology and
Special Education

1993

ABSTRACT

ASSESSING THE COMPETENCY OF TEACHERS, CURRICULUM SPECIALISTS, AND PROSPECTIVE TEACHERS IN EDUCATIONAL MEASUREMENT IN BAHRAIN

By

Rashid Hammad Aldosary

Educators in general and teachers in particular use classroom tests to make a variety of decisions.

This study was conducted to assess the competency of Bahrain teachers, curriculum specialists, and prospective teachers in educational measurement, to detect their strengths and weaknesses in specific areas in educational measurement (such as item analysis and score interpretation) and to investigate the relationship between competency in educational measurement and classroom testing activities, measurement preparation, and background variables.

Two research instruments were used in this study: a measurement test (65 multiple-choice items) and a questionnaire (23 items). The two research instruments were administered to 1300 in-service teachers, 146 curriculum specialists, and 221 prospective teachers. The response rate was 82 percent for the teachers, 67 percent for the curriculum specialists, and 52 percent for the prospective teachers. The researcher was unable to generalize to the population of prospective teachers.

The three groups in the study had a reasonable amount of measurement and statistics coursework, with more measurement coursework than statistics.

About one third of the teachers had never been exposed to statistics. All three groups need more training in measurement and statistics. The subjects, in general, demonstrated about average performance on the educational measurement test, except for the teachers who scored below average. The mean test score was 26.72 (41 percent) for the teachers, 37.49 (58 percent) for the curriculum specialists, and 34.07 (52 percent) for the prospective teachers. The scores on the measurement test ranged from 9 to 48 for the teachers, from 19 to 50 for the curriculum specialists, and from 18 to 49 for the prospective teachers. The results indicated that the three groups need training in measurement in general, and especially the areas in which they were weak, such as "item analysis, "score interpretation", and "standard error of measurement".

About half of the teachers spend more than 20 percent of their professional time on classroom testing activities. The most preferred item type for the teachers and prospective teachers was the essay, whereas the completion type was the most preferred item type for the curriculum specialists.

Correlational results revealed that there was a significant relationship ($P<.05$) between competency in educational measurement and several variables such as purposes of using classroom tests, developing a test plan, conducting item analysis, and self-assessment of the general competency in measurement. Yet the proportion of variance in the measurement test accounted for by each of these variables was small or negligible.

ANOVA results showed that there was a significant difference in measurement competency among the three groups. Also, female teachers were more competent in measurement than male teachers. Those who had measurement and statistics coursework manifested better performance on the measurement test than those with no measurement and statistics at all (except for the prospective teachers with respect to statistics). These results also apply to the

test plan and item analysis self-ratings, but for the teachers only. The gender of respondent had an effect on the competency in educational measurement from one group to another, but age has no bearing on the competency of respondents from one group to another.

Multiple regression analysis indicated that statistics coursework was among those variables that best predict competency in measurement.

Further training in educational measurement for all three groups is highly recommended.

If there is an opportunity to learn, take it.
If there is no opportunity to learn, make it.

— Rashid Aldosary

ACKNOWLEDGMENTS

First, I would like to express my thanks and gratitude to the Bahrain Ministry of Education for granting me a scholarship to pursue my Ph.D. program in Measurement, Evaluation, and Research Design at Michigan State University.

Thanks and deep appreciation are due to Dr. Irvin Lehmann, my adviser and chairperson of my doctoral committee, for his cooperation, patience, understanding, and valuable suggestions regarding the research instrument.

I am grateful to Dr. William Mehrens for being a cooperative member in my doctoral committee, and for his suggestions during the preparation of the test instrument. Thanks go to Dr. Peggy Riethmiller and Dr. Antonio Nunez for their time and understanding, and for being cooperative members in my doctoral committee.

My sincere thanks and appreciation to Dr. Malik Balla in the Department of Arabic at Michigan State University, Ivan Filmer, Ahmad Y. Alghamdi, Naji Alarfaj, Humood Almuqate, and Mohammad Al-Ameeri for their careful review of the research instrument, and for their valuable suggestions.

I am really indebted to Yousif Alzayani, former Registrar of the University of Bahrain, and Khalid Bumtea, Coordinator of Registration at the University of Bahrain for their great cooperation before and throughout the data collection process.

Thanks are due to Dr. Abbas Adebee, Dean of the College of Education, and Dr. Hussain Alsadeh, Head of the Department of Education at the University of Bahrain, for facilitating the data collection process.

Thank you to Dr. Ibrahim Yousif Alabdallah, Director of Curricula in the Bahrain Ministry of Education, for his cooperation during the data collection process.

Great thanks and sincere appreciation go to all curriculum specialists, teachers, prospective teachers, and professors of education at the University of Bahrain for their cooperation and for giving generously of their time.

Finally, my deepest and sincerest love and appreciation are due to my wife and best friend, Noora, for her understanding, sacrifice, and relentless support.

TABLE OF CONTENTS

Page

LIST OF TABLES. x

CHAPTER

I. THE STATEMENT OF THE PROBLEM 1

Introduction 1
Purpose of the Study 6
Importance of the Study 7
Research Questions 12
Research Hypotheses 12
Limitations 14
Definition of Terms 15

II. REVIEW OF RELATED LITERATURE 16

Introduction 16
Integrating Testing With Teaching 16
Teachers' Classroom Testing Practices 19
Teacher Training in Educational Measurement 23
Teachers' Competency in Educational Measurement 30
Educational Measurement Needs of Teachers
Prospective Teachers, Curriculum Specialists, and
Other Supervisory Personnel 34
Conclusions 42

III. RESEARCH DESIGN AND PROCEDURE 45

Introduction 45
Research Instruments 45
Subjects and Sampling Procedures 47
Characteristics of the Subjects 48
Data Collection 56
Follow-up of Non-Respondents 57
Data Analysis 57

Summary 58

IV. ANALYSIS AND INTERPRETATION OF THE DATA 59

Introduction 59
Classroom Testing Practices and Measurement preparation 59
Performance of the Subjects on the TEST 68
Testing the Research Hypotheses 71
Multiple Regression Analysis 89
Summary 90

V. SUMMARY, DISCUSSION, CONCLUSIONS, AND SUGGESTIONS. 93

Summary of Purposes and Procedures 93
Discussion and Conclusions 94
Recommendation 99
Suggestions for Further Research... 100
Reflections 102

APPENDICES
A. THE ENGLISH VERSION OF THE TEST AND THE QUESTIONNAIRE 104
B. THE TABLE OF SPECIFICATIONS OF THE TEST 122
C. THE TEST ANSWER KEY 123
D. THE ARABIC VERSION OF THE TEST AND THE QUESTIONNAIRE 124
E. CORRESPONDENCE 141
F. TABLES 4.14 TO 4.16 154

BIBLIOGRAPHY 157

LIST OF TABLES

Table No.: Page

3.1 Information on the Sample of Teachers by Sex and grade Level 47
3.2 Information on the Participating Schools by Sex and Grade Level 48

3.3 Information on the Population of Curriculum Specialists and Prospective Teachers Used in the Study 48
3.4 Information on the Response Rate for the Total Sample and for each Group..... 49
3.5 Distribution of the Respondents' Age in the Study 49
3.6 Distribution of the Respondents' Sex in the Study 49
3.7 Distribution of the Respondents by Highest Degree Held 50
3.8 Information on the Highest Degree Held Obtained from a College of Education 51
3.9 Distribution of the Respondents Whose Major was Education in the Bachelors Program 51
3.10 Distribution of the Respondents by Major in College 52
3.11 Distribution of the Respondents by Teaching/ Supervision Subject 54
3.12 Distribution of the Respondents by Grade Level 55
3.13 Distribution of the Respondents by the Average Number of Years in Teaching 55
4.1 Distribution of the Respondents by Measurement Coursework Completed 61

4.2 Distribution of the Respondents by Statistics Coursework Completed 61
4.3 Distribution of the Respondents by percentage of Time Devoted to Classroom Testing 62
4.4 Distribution of the Respondents by Purposes of Classroom Test Use 63
4.5 Item Types Used in Classroom (percentages) 64
4.6 Item Types Ranked According to Preference 65
4.7 Distribution of the Respondents by Test Plan Preparation.. 65
4.8 Distribution of the Respondents by Item Analysis Practice 66

4.8 Distribution of the Respondents by Item Analysis Practice 66
4.9 Distribution of the Respondents' Self-Ratings Regarding Their General Competency in Educational Measurement 66
4.10 Distribution of the Respondents' Self-Ratings Concerning Their Competency in Test Construction 67
4.11 Distribution of the Respondents' Practical Training Needs in Educational Measurement (Percentages) 67
4.12 KR20 Reliability and SEM Estimates for the TEST 68
4.13 Distribution of the Respondents' Performance on the TEST by Content Area 70
4.14 Item Analysis Data for the TEST: Teachers 154
4.15 Item Analysis Data for the TEST: Curriculum Specialists 155
4.16 Item Analysis Data for the TEST: Prospective Teachers 156
4.17 Correlations Between the Scores on the TEST and Classroom Testing Activities Variables, and Measurement Preparation Variables 73
4.18 ANOVA for the TEST Scores With "Purposes of Classroom Test Use" Variable for the Teachers 74
4.19 ANOVA for the TEST Scores with the "Perceived Level of General Competency in Measurement" for the Teachers 75
4.20a Distribution of the Respondents' Mean Performance on theTEST 77
4.20b ANOVA for the TEST Scores with the "Type of Respondent" 77
4.21a Distribution of the Respondents by Measurement Coursework Completed 78
4.21b ANOVA for the TEST Scores with "Measurement Coursework" 79
4.22a Distribution of the Respondents by Statistics Coursework Completed 80
4.22b ANOVA for the TEST Scores with "Statistics Coursework" 81
4.23a Distribution of the Respondents by Test Plan Preparation 83
4.23b ANOVA for the TEST Scores with "Test Plan" 82
4.24a Distribution of the Respondents by Item Analysis Practice 83
4.24b ANOVA for the TEST Scores with "Item Analysis" 84
4.25 ANOVA for the Joint Effects of Gender and Type of Respondent with the TEST Scores 85
4.26a Distribution of the Respondents' Age in the Study 86
4.26b ANOVA for the Joint Effects of Age and Type of Respondent with the TEST Scores 86
4.27 t-test for the Difference Between Male and Female Teachers in Educational Measurement 87
4.28 ANOVA for the Grade Level of the Teacher with the TEST Scores 88
4.29a Information on the Sample of Teachers by Sex and Grade Level 89
4.29b ANOVA for the Joint Effects of Teacher's Gender and Grade Level with the Competency in Educational Measurement 88
4.30 Multiple Regression Analysis for the TEST Scores with Measurement Preparation, and Background Variables For the Teachers 90
4.31 Summary of the Major Findings in the Study 92

CHAPTER I
THE STATEMENT OF THE PROBLEM

Introduction

It seems reasonable to believe that the relationship between measurement and instruction is inextricable. Rudman et al. (1980, p.2) stated that "there is a growing sense of need today to link assessment with teaching and to make instructional decisions more rational and less intuitive".

Mehrens and Lehmann (1991, p.vii) reported that "educators have always been concerned with measuring and evaluating the progress of their students", and that "the role of measurement is to provide decision makers with accurate and relevant information" (p.3.). Hence, the linkage between testing and instruction is a fundamental concern in educational practice (Stiggins & Bridgeford, 1986), and all educators know that testing is a regular part of the school routine (Gullickson, 1984).

Rudman (1989, p.2) indicated that

> Testing and teaching are not separate entities. Teaching has always been a process of helping others to discover new ideas and new ways of organizing that which they learned, whether this process took place through systematic teaching and testing, or whether it was through a discovery approach, testing was, and remains, an integral part of teaching.

Rudman, among others, says that testing can be used to help teachers and administrators make promotion and retention decisions, and can be a useful tool for measuring the effectiveness of teaching and learning.

Stiggins and Conklin (1988, p.3) illustrated the linkage between classroom instruction and teacher's assessment of student achievement and reported that "the quality of classroom instruction is a function of the teacher's assessment of student achievement". They state that sound, daily assessment of student learning is important for several purposes, such as diagnosing student needs, assigning grades, and evaluating the impact of instruction. In addition, McMorris and Boothroyd (1992) reported that 69 percent of the teachers they interviewed (n=42) stated that their main reason for classroom testing is to assess students' mastery and understanding of the material taught.

Research on teachers' testing practices has consistently revealed that tests are used extensively in the classroom (Green & Stager, 1986-1987; Stiggins, 1991), and that "assessing, grading, and evaluating students has been identified as one of a teacher's six core job functions" (Schafer, 1991, p.3). In addition, Stiggins & Conklin (1988), Schafer (1991), & Stiggins (1991) found that teachers do spend from a quarter to one third of their instructional time on assessment-related activities. Based on their research on testing practices, Green and Stager (1986-1987) found that teachers spend 10 to 15 percent of their time in testing, and that 40 to 50 percent of students' course grades are based on test scores. As Mehrens and Lehmann (1984) have indicated, "teacher-made tests are frequently the major basis for evaluating students' progress in school. One would have

great difficulty in conceptualizing an educational system where the child is not exposed to teacher-made tests" (p.56). A major part of the teacher's preparation programs in educational measurement and classroom testing should be devoted to training them to meet the classroom assessment demands (Stiggins & Conklin, 1988).

Newman and Stallings (1982) noted that over several decades, educators have been aware of the need for teacher competency in measurement, and that "much of the literature on this subject points to a lack of emphasis on measurement training. Measurement educators almost unanimously agree that this adversely affects classroom practices...but little is known about teacher competency in classroom testing" (p.3).

Mayo (1964) surveyed what he referred to as "experts" (teachers, principals and superintendents, college and university professors, and research specialists) to judge what beginning teachers should know about measurement. For example, he found that knowledge of general principles of test construction, and the ability to interpret achievement test scores were rated very important. In another study regarding students' first course in measurement, Mayo (1970) found that measured competency in measurement was mediocre not only among graduating seniors in teacher training, but also among the same persons two years after graduation. Mayo's (1967) landmark study, titled *Preservice Preparation of Teachers in Educational Measurement*, revealed that test scores on the measurement competency test (MCT) were related to teaching field, and amount of testing and measurement coursework taken. He concluded that, in general, there is agreement on the importance of some measurement competencies for teachers, and that beginning teachers do not demonstrate a very high level of measurement competency. Based on the findings of his study, Mayo

recommended that some measurement coursework be mandatory, yet meaningful, and that further research on teacher's competency in measurement and testing be conducted. Green and Stager (1986-1987) reported that there has been little change in teacher's competency in measurement and testing since Mayo's (1967) landmark study, and that there is a lack of competence in testing techniques in general among teachers.

Farr and Griffin (1973, p.19) made a case in support of two hypotheses: "the first is that in general teachers have quite limited knowledge of measurement concepts. The second is that teachers are not being taught what they need to know about measurement in order to be more effective teachers". Newman and Stallings (1982, p. 12) also indicated that "measurement training practices do not seem to be any more effective now than they were over a decade and a half ago". Similarly, Gullickson (1986, p.347) noted that "today's measurement and evaluation textbooks still strongly reflect the stated emphasis" from Mayo's studies (1964, 1967). Gullickson (1986) concludes that teachers would improve their instruction if they were more knowledgeable about measurement and utilized better measurement techniques. He adds that educators need to determine which measurement tools and strategies are practical and beneficial for teachers.

Recent research on classroom assessment, teacher-made tests, and teacher's knowledge and competency in measurement (Stiggins, 1985, 1988, 1991; Stiggins & Bridgeford, 1985; Marso, 1985; Marso & Pigge, 1987, 1989, 1991; and Carter, 1984) revealed several important findings that have implications for research on, and teacher's preservice and in-service training and competency in measurement and classroom testing. First, public attention and a large and growing body of research have

concentrated on standardized tests and testing issues, whereas research on testing practices and teacher-made tests have received little emphasis or have largely been neglected despite the fact that extensive literature exists on classroom testing and evaluation practices. Second, the advice usually given to preservice and in-service teachers about classroom test development and test use relies heavily on professional judgment. Third, the content of teacher preservice measurement and testing courses mainly designed to facilitate the effective development and use of teacher-made tests may not reflect the actual needs of classroom teachers. Fourth, a thorough and comprehensive reexamination of preservice measurement courses and in-service testing activities is needed. Finally, construction of teacher-made tests, and specific test item writing skills need more emphasis in in-service teacher training programs.

Supporting Mayo (1967), Newman (1981, p.2) states that "if teacher preparation institutions have acted upon Mayo's recommendations to improve measurement, it would be expected that today's teachers, especially young teachers and those with more recent degrees, would be adequately prepared and competent in classroom testing principles and techniques". Newman continues that "there is an urgent need to study and document teacher competency in classroom testing" (p.23).

Nitko (1991) indicates that much work still needs to be done in developing teacher's assessment competencies, and makes it clear that "reformers' call for increased relevance of course content will be ineffective if preservice and in- service teachers do not enter testing and measurement classrooms" (p.2).

Purpose of the Study

The study was conducted for the following purposes:

1. determining how well in-service teachers, curriculum specialists, and prospective teachers in Bahrain understand measurement principles and the techniques of test construction, and in which areas these three groups are most and least proficient.
2. determining how in-service teachers, curriculum specialists, and prospective teachers in Bahrain have been prepared in measurement, as well as in basic statistics related to measurement.
3. collecting information from in-service teachers, curriculum specialists, and prospective teachers in Bahrain regarding the percentage of their professional time they devote to testing, the purpose for which they use their self-constructed tests, the item types they write for their self-constructed tests, and the item types they rank order in terms of their preference.
4. collecting information on whether these three distinct groups develop a test plan (e.g., table of specifications) prior to writing test items, and conduct an item analysis. In addition, information was collected on how well the subjects in each group assess their level of competence in measurement and testing in general, and test construction in particular, and their practical needs in measurement.
5. investigating the nature of the differences among the three groups concerning their competency in measurement, and whether these differences are related to sex, age, grade level, amount of measurement coursework, recency of measurement coursework (the same applies for basic statistics), highest degree held, and recency of highest degree held.

6. investigating the nature of the differences between males and females, for each group separately, regarding their competency in measurement.
7. investigating the nature of the relationships between the three groups' competency in measurement and such variables as: amount of measurement coursework completed, amount of statistics coursework completed, classroom testing practices, years of experience in teaching/supervision, grade level of teaching/supervision, and subject of teaching/supervision.

Importance of the Study

The significance of this study is that it provides current information on an essential, important and often less emphasized and less researched dimension of teaching; namely, knowledge and understanding of measurement principles. Moreover, conducting such a study in Bahrain, where there is no empirical information on the current level of knowledge and competency in measurement among in-service teachers, curriculum specialists, and prospective teachers, should provide valuable information for educators and researchers in Bahrain, and valuable comparative information for researchers outside Bahrain.

Educators in Bahrain are in need of such a study to help them make less subjective decisions. As Alabdallah (1990) noted, most teachers in Bahrain are deficient in diagnosing strengths and weaknesses in student academic achievement and learning, which led them to incorrectly employ classroom assessment. Moreover, more than 85 percent of teachers in Bahrain have some deficiency in constructing and correctly employing

classroom test items that are valid, that require different levels of difficulty, and require functioning at higher levels, such as analysis and synthesis. Alabdallah continues that the implementation of observation techniques and other measurement devices is limited among teachers and is not flawless. More than 90 percent of in-service teachers, and even the new graduates and those with a high diploma in education (one year of full-time study beyond the bachelors degree), according to Alabdallah, are not proficient in using basic statistical methods that deal with test reliability and validity in particular, and other measurement topics in general.

Based on their research findings on the teacher-made tests they analyzed, Marso and Pigge (1991) suggest more attention and emphasis on test construction and item writing skills in teacher preservice training and in measurement courses. They add that teachers should be encouraged to develop the necessary skills in test construction, and use these skills to construct test items that function at higher cognitive levels.

Research conducted by Marso and Pigge (1987, October) on teacher-made tests and testing has revealed that about 80 percent (n=328) of the teachers reported that they rarely or never calculate test means or standard deviations, and that more than 50 percent of them reported never having estimated test reliability or conducted item analysis.

Stiggins, Bridgeford, and Conklin (1988, pp. 14-18) concluded that

> ...teacher training only in paper and pencil measurement methods face real difficulties in the classroom assessment environment.... There are fundamental, far reaching differences between the science of testing and the assessment demands of the classroom. We have been aware of these differences for decades and have

> failed to address them... The result of poor measurement is poor decision making. Perhaps the most compelling of our failure to address classroom assessment needs will be the continued alienation of teachers from systematic assessment and evaluation processes.

Not only is it important to assess the competency of teachers and prospective teachers in measurement, it is equally important to assess curriculum specialists in Bahrain because there is no empirical information concerning their competency in measurement, and because they play a leading role in Bahrain public schools. Curriculum specialists are the link between the Department of Curricula at the Bahrain Ministry of Education and the public schools. Their main tasks include participating in the in-service training of teachers, contributing to curriculum development and evaluation, and collaborating with public school teachers to improve teachers' instruction. Moreover, they conduct research and prepare workshops on instructional methods and assessment techniques and follow-up these studies and workshops. Finally, they supervise what is referred to in Bahrain as a "subject team" of public school teachers who develop instructional methods, design appropriate assessment techniques, construct mid- semester and final examinations, design diagnostic materials for "slow-learning" students, analyze school syllabi, study and analyze initiatives submitted by teachers to develop classroom assessment and instruction, and construct national examinations in the fall and spring semesters for the 9th and 12th grades and construct "model" tests for other grades to guide instruction (Alabdallah, 1989). Given these responsibilities, it is essential to assess the competency of curriculum specialists in measurement.

Research on measurement competency and needs (e.g., Marso and Pigge, 1987, March, 1988, November, 1989, March) has concentrated on curriculum specialists' perceptions concerning teachers' needs in measurement, and how curriculum specialists can assist teachers in identifying and alleviating the most common errors in teacher- made tests. These studies, however, have never attempted to question the competency and investigate the needs of curriculum specialists in measurement.

Schafer highlights this issue as he states that

> Administrators and other supervisory personnel in schools also use assessment, yet measurement education in these programs has been found to be rarer than it is in teacher education programs. If such training is important for teachers it is also important for principals, curriculum specialists, etc., who provide instructional leadership. Moreover, supervisory personnel often have responsibilities regarding evaluative judgments about educational programs for which a broader methodology, such as statistical decision-making, is pertinent. It would be useful to expand the dialogue about teacher education in measurement to include the specific assessment skills needed by those in supervisory roles (Schafer, 1991, p.6).

It is hoped that the information provided by this study will

1. shed light on the actual preparation of teachers, curriculum specialists, and prospective teachers in Bahrain in measurement, and reveal each

group's strengths and weaknesses in these areas in order to plan measurement programs for them as needed.

2. investigate whether incompetence in measurement is a major problem within each group independently, or the major problem among the three groups in general.

3. reveal the differences among teachers, curriculum specialists, and prospective teachers in Bahrain in measurement in order to help educational policy makers in Bahrain reexamine and reevaluate the quality of preservice and in-service training programs in measurement.

4. guide the educational policy makers and administrators at the Bahrain Ministry of Education to give serious consideration to the training of curriculum specialists in measurement, and to reevaluate the criteria by which these curriculum specialists are usually selected for their prospective positions in the curriculum department.

5. establish a vital link between the Bahrain Ministry of Education and the University of Bahrain to design a long-range plan, based on empirical information, for preservice and in-service training in measurement for teachers, curriculum specialists, and prospective teachers.

6. lead to more research on educational measurement needs in Bahrain, and in the other Arabian gulf states that are currently on their way to unify public school curricula in the whole region.

7. open a new horizon and encourage more research on the measurement needs of curriculum specialists.

Research Questions

This study has attempted to answer the following questions.

1. What is the actual preparation of Bahrain teachers, curriculum specialists, and prospective teachers in measurement?
2. What are the current testing practices of Bahrain teachers, curriculum specialists, and prospective teachers?
3. In what areas of educational measurement do curriculum specialists, teachers and prospective teachers need practical training?
4. What is the current level of knowledge and understanding of Bahrain teachers, curriculum specialists, and prospective teachers in measurement?
    a. How competent is each group in measurement?
    b. In which measurement area(s) is each group the most and the least competent?

Research Hypotheses

1. For each of Bahrain teachers, curriculum specialists, and prospective teachers there is no significant relationship between their measurement competency and
    a. highest degree held.
    b. recency of highest degree held.
    c. major subject of specialization.
    d. major subject of teaching/ supervision.
    e. grade level of teaching/ supervision.
    f. years of experience in teaching.
    g. years of experience in supervision.

h. amount of measurement coursework completed.

i. recency of measurement coursework completed.

j. amount of statistics coursework completed.

k. recency of statistics coursework completed.

l. percentage of professional time devoted to test construction.

m. purposes of using teacher-made classroom tests.

2. For each of Bahrain teachers, curriculum specialists, and prospective teachers there is no significant relationship between their measurement competency and

a. developing a test plan prior to writing test items.

b. conducting item analysis on self-constructed classroom tests.

c. self-assessment of the level of competency in measurement in general.

d. self-assessment of the level of competency in test construction principles.

3. There is no significant difference in measurement competency among Bahrain teachers, curriculum specialists, and prospective teachers.

4. There is no significant difference in measurement competency (for each group separately) between those who had measurement coursework and those who did not.

5. There is no significant difference in measurement competency (for each group separately) between those who had coursework in statistics and those who did not.

6. There is no significant difference in measurement competency (for each group separately) between those who develop a test plan and those who do not.

7. There is no significant difference in measurement competency (for each group separately) between those who conduct item analysis and those who do not.

8. There is no gender by type of respondent (e.g., a teacher) interaction regarding their competency in educational measurement.

9. There is no type of respondent (e.g., a teacher) by respondents' age interaction regarding their competency in educational measurement.

10. There is no significant difference between male and female teachers in their measurement competency.

11. There is no significant difference among teachers at different grade levels in measurement competency.

12. There is no gender by grade level interaction for the teachers concerning their measurement competency.

Limitations

This study is limited to the populations of in-service teachers, curriculum specialists, and prospective teachers in Bahrain. No generalizations regarding the results of this study are expected to be made beyond these three populations, nor are any inferences to be drawn concerning any other educational populations related to these three groups (such as public school principals).

Definition of Terms

The following terms are defined in the context in which they were used throughout this study.

ASSESSMENT: "The process by which one person attempts to find out about the knowledge, attitudes, or skills possessed by another" (Rowntree, 1981, p.14).

COMPETENCY: "Any single knowledge, skill,... the possession of which is believed to be relevant to the successful practice of teaching" (Mehrens, 1987, p.197).

TEACHERS: All in-service public school teachers in Bahrain.

CURRICULUM SPECIALISTS: All in-service subject-area curriculum specialists in the Department of Curricula; Bahrain Ministry of Education.

PROSPECTIVE TEACHERS: All preservice teachers in Bahrain who are currently in the final year of their undergraduate program at the University of Bahrain.

CHAPTER II

REVIEW OF RELATED LITERATURE

Introduction

This chapter is divided into five major sections. The first section deals with the relationship between testing and teaching, and shows how they are integrated. The second section emphasizes teachers' classroom testing practices and highlights the research findings about these practices. Teacher education and training in educational measurement constitutes the central theme of the third section along with the related shortcomings in that education and training. The fourth section discusses the past and present research findings on teachers' competency in educational measurement. The fifth section discusses what teachers, prospective teachers, curriculum specialists, and supervisory personnel need in educational measurement, either for their classroom assessment practices or for administrative purposes, based on research findings and the suggestions of measurement experts in the field. Some overall conclusions will follow.

Integrating Testing with Teaching

There is no doubt that teaching is a decision-making process. Mehrens and Lehmann (1991) state that "... educational decisions are continually being made. These decisions should be based on information

that is accurate. The responsibility of gathering, using, and imparting that information belongs to educators" (p.vii). Therefore, "good schooling depends on good decisions, and good decisions depend on good information" (Mehrens & Lehmann, 1987, p.36). Hence, one would view testing as systematic procedures for purposes of observing and classifying students' performance in order to obtain instructionally-relevant information for sound decision making (Nitko, 1989).

By and large, testing and teaching can be viewed as related entities. Teaching can help us discover new ideas and organize them through several ways. Thus, testing can be a useful tool by which teachers can gain a general idea of what their students bring to instruction and modify their instructional strategies if needed. Testing can also help teachers and administrators make promotion and retention decisions, and can be a useful tool for measuring the effectiveness of teaching and learning (Rudman, 1989). Further, Ward stated that "to teachers, testing should be an integral part of the instructional programs, and there should be a high match between tests and the curriculum" (1980, p.12).

Mehrens and Lehmann (1987) reported that teachers basically use data from their self-produced classroom tests and other sources, such as observations, to serve the main functions of teaching, such as setting realistic goals for students. Nevertheless, it is really surprising to know that "... most teachers separate testing from teaching" (Rudman, 1987a, p.6). Moreover, classroom assessment can encourage good study habits by students' frequent review of materials, and can provide students with feedback regarding their strengths and weaknesses (Mehrens & Lehmann, 1987).

Educational measurement and testing can play several other leading roles in instruction, such as selecting learning experiences and procedures of instruction, organizing learning experiences, and aiding in the supervision and administration of instruction. They are valuable for the continuing development of teachers and coordination of instructional efforts. Educational measurement is conceived, then, as an integral part of instruction, not as separate from instruction (Tyler, 1951), because it is impossible to imagine how teachers can diagnose students, follow up student learning, and communicate with parents without daily classroom assessment of student achievement (Stiggins & Conklin, 1988).

Educators should carefully and consciously attempt not only to help teachers make decisions but also help teachers understand how to select the relevant and accurate information that they need, along with their intuition and judgment, to make correct decisions. Teachers should be knowledgeable about the quality of information they gather and use, and the criteria by which they judge that information. Underlying all of this discussion is the fact that most of the information that teachers use for making decisions about student learning and achievement comes from a variety of sources. This can include informal or systematic observations of students, teacher-made tests, standardized tests, students' oral presentations, and completed projects. Hence, teachers and administrators must understand the important potential role educational measurement and evaluation play in integrating testing with teaching. Adopting this perspective should lead us to understand that our knowledge of the relationship between assessment and instruction will alleviate the persistent notion that testing takes time away from instruction (Rudman, 1987a).

Finally, we believe, as do Mehrens and Lehmann, that

> ...unless there is administrative support for classroom assessment, until administrators and teachers come to appreciate the role and value of evaluation in the teaching-learning process, and until through our teacher-training institutions or through high quality professional development programs we prepare both teachers and administrators to develop and interpret the results of classroom assessment, our pupils will not be receiving maximal instruction (Mehrens & Lehmann, 1987, p.42).

Teachers' Classroom Testing Practices

As noted earlier, a number of researchers who have examined teachers' classroom testing practices have documented the fact that teachers spend between ten and thirty-three percent of their instructional time on testing and assessment - related activities (Stiggins & Conklin, 1988; Stiggins, 1991; Schafer, 1991; Green & Stager, 1986-1987; & Gullickson, 1982). In addition, Green and Stager (1986-1987) found that 40 percent to 50 percent of students' course grades are based on test scores. Further, many teachers place more weight on their own tests than on anything else when assigning students' final grades (Boothroyd, 1990).

Fleming and Chambers (1983) reviewed 342 twelfth-grade teacher-made tests and found that, by and large, teachers tend to omit test directions, use illegible test copies, and omit the point values assigned each test item. Fleming and Chambers concluded that their findings suggest that teachers may not view their own tests "as a means for quantifying students'

performance" (p.36). This appears to confirm previous studies according to Fleming and Chambers.

Harris (1973) surveyed 145 NCME members to determine the extent of their agreement on 40 controversial issues in educational measurement. Regarding those issues pertinent to classroom teachers' testing practices, Harris found that sixty-nine percent of the members surveyed agreed that the majority of teachers rely heavily on their subjective judgment in evaluating and grading their students, and that seventy-six percent of those surveyed agreed that most classroom teachers use very few statistical techniques in evaluating and grading their students.

Marso and Pigge (1989, 1991) analyzed a sample (n=175) of teacher-made tests to determine item types frequently used, item cognitive functioning levels, and the frequency and nature of test construction errors. They found that the most frequently used item types are matching, multiple-choice, and completion. Furthermore, they found that teacher-made test items primarily function at the knowledge level of the cognitive domain (according to Bloom's taxonomy), and that the matching test items were "the most construction error prone followed by completion, essay, and true false" (Marso & Pigge, 1991, p.283). In a related study on classroom teachers' testing practices, testing proficiencies, and testing resources, Pigge and Marso (1988) surveyed 586 school supervisors and principals, and 326 school teachers regarding how supervisors can help teachers identify and alleviate the most common test construction errors in teacher-made tests. They reported that the "average" teacher gives 54.1 formal tests during the school year, and that the "average" teacher-made test contains 37.9 items. In addition, 50 percent of the teachers said they construct 75 percent or more of their classroom test items. Pigge and Marso also found that 72

percent of teacher- made test items function at the knowledge level, 11 percent at the comprehension level, 15 percent at the application level, and about 1 percent measure higher- order thinking skills (e.g., analysis, synthesis, and evaluation). These findings are similar to those results reported by Gullickson (1982).

It is worth noting that teachers tend to prefer certain types of items for their self-produced classroom tests. Newman and Stallings (1982) found that teachers' rankings of item types from most to least in terms of preference were "(1) completion, (2) multiple-choice, (3) matching, (4) true-false, (5) short answer, (6) calculations, and (7) essay" (p.11). Newman and Stallings also noted that these rankings differed from those obtained by Reynolds and Menard (1980) although the teachers' rank order of multiple-choice items (ranked 2) did not change.

One will recognize that classroom testing practices and the contents of teacher -made tests in other countries such as Japan may not be much different from those in the United States. Yamagishi (1991) reports that, based on a survey of schools in 10 regions in Japan, Japanese teachers mostly favor those tests that are designed to accompany classroom textbooks. Japanese teachers' main reason for using these "text" tests, according to Yamagishi, is due to the lack of time available to construct their own classroom tests. Detailed analysis of these tests revealed that they were poorly written, emphasized rote memory, and tested fragmented knowledge. Yamagishi found that short answer and essay items are rarely found in the Japanese teacher-made tests in language arts. In social studies, she found that 80 percent of the items were of the recall or recognition type, and 90 percent of them dealt with facts. In mathematics, Yamagishi reports that the majority of the items measured computation and

understanding, whereas 5.1 percent were at the application level, and 3.3 percent functioned at the analysis level.

It seems that recent research on classroom testing practices verifies the research findings of Gullickson (1982), who noted that short answer and matching item types are the most popular among teachers and both types are restricted to lower cognitive level (e.g., knowledge), despite the fact that this contradicts what Yamagishi has reported about teacher-made tests in Japan with respect to the use of short answer items in language arts. Gullickson's results suggest that neither the cognitive level of the items nor item difficulty were adequately addressed. If teacher-made tests are targeted toward the lower cognitive level, and students' grades are based on their achievement at this level, students' motivation will be focused on lower cognitive level learning. Gullickson concluded that there are three important factors that influence classroom testing practices: expertise, time, and tools available for the teachers' use.

Further examination and analysis of the results of the teachers' classroom testing practices suggest that there are two main factors that contribute to the frequency of use of teacher-made tests: subject matter and grade level. Boothroyd (1990), Stiggins and Bridgeford (1985), and Green and Stager (1986-1987) reported that the use of teacher-made tests varies from subject to subject. For example, Stiggins and Bridgeford (1985) noted that teachers are more likely to use teacher-made tests for assessing student competency in mathematics (86 percent) or science (88 percent) as compared with speaking (71 percent) and writing (74 percent). Stiggins and Bridgeford state that, in terms of grade level, the number of teacher-made tests increases with grade level, and that a positive relationship has been detected between grade level and the proportion of teachers who construct

their own classroom tests. They also report that 68 percent of the second grade teachers surveyed said they develop their own classroom tests compared to 91 percent of the eleventh-grade teachers. Furthermore, Marso and Pigge's (1988) survey of teachers (n=326) revealed that secondary school teachers test more frequently than elementary school teachers.

Teacher Training in Educational Measurement

"Historically, teachers have not been required to complete coursework in educational measurement to attain state teacher certification" (Boothroyd, 1990, p.19). Noll (1955) reported that only five states required completion of one educational measurement course for obtaining state teacher certification. Based on his survey of 80 teacher-education institutions, Noll found that 82.5 percent of them offered courses in educational measurement, but only 21.2 percent of these institutions that offer a measurement course required that course for undergraduates, and only 13.6 percent of them required it for all undergraduates preparing to be teachers.

In a study regarding the experience of teachers with tests and formal training in measurement, Goslin (1967) found that more than 20 percent of secondary school teachers surveyed (n=301) had never been exposed to a course in tests and measurement. He also found that 39.2 percent of secondary school teachers in his sample have low familiarity with tests. Goslin further reported that more than 40 percent (N=1440) of all teachers surveyed had only one course in measurement. Goslin concluded that his study indicated that "teacher involvement in testing...is of potentially great influence on the educational system and the children in it" (p.129).

In his landmark study, Mayo (1967), surveyed 2,877 seniors in 86 randomly selected teacher-training institutions concerning their training in educational measurement. Mayo found that only 30 percent of the student teachers surveyed reported having taken a course in educational measurement. He also concluded that some teachers, especially elementary teachers, have "a strong bias against statistics, apparently because they see no relation to their work" (p.56).

Based on his nationwide survey of 896 accredited institutions offering teacher education programs, Roeder (1972, 1973) indicated that 57.7 percent of the institutions surveyed "did not require their prospective elementary teachers to complete a course in evaluation" (p.239). Roeder comments that

> when one considers the increasing importance which report cards marks and test scores play in the lives of our children, it is readily apparent that the failure of teacher education institutions to provide prospective teachers with at least a minimal understanding of evaluation is inexcusable" (p.143).

Ward (1980, 1982) indicated that while many teachers deal daily with testing- related issues and activities, the vast majority of them have had no formal training in educational measurement. The results of Ward's survey on teachers' attitudes toward and knowledge of testing (n=209) revealed that 29 percent of the surveyed teachers indicated they had never taken any measurement courses; 29 percent had taken one; and 42 percent of them had taken two. Moreover, 70 percent of the teachers reported that they had never taken any in-service training in educational measurement.

This lack of training or adequate training in educational measurement seems to have an adverse impact on instructional practices and student assessment in the classroom. In this regard, Green (1963) states that

> yet it is not unusual to find a teacher who has had no specific training in measurement and evaluation and who has evolved his testing procedures through trial and error. Such a teacher often designs tests which measure objectives other than those which have been emphasized in his classroom. Because his examinations emphasize trivial points which had been either neglected or given cursory considerations, he has little notion of the effectiveness of his instructional objectives. For this teacher, tests may become a series of puzzles designed to fool the student instead of instruments for measuring the relative attainment of classroom objectives (pp. 1 -2).

Schafer and Lissitz's (1987) review of literature concerning measurement training for school personnel appears to be comparable with previously discussed research findings. They noted that "a significant proportion of school personnel do not receive much training in measurement methods" (p.61). They conclude that teachers' performance on tests of measurement principles, and analysis of teacher-made tests, revealed that teachers lack training to utilize measurement practices.

Recent research on measurement-related issues (Stiggins, Griswold, & Frisbie, 1986; Stiggins, Frisbie & Griswold, 1989) has shown that there is a lack of teacher training in grading. Stiggins et al. studied 15

high school teachers, utilizing what they refer to as "intensive case study methodology". The researchers prepared and used a comprehensive framework of 34 grading issues to serve as a basis for observing teachers' grading practices. The two major issues that the researchers explored were: (1) the nature and technical quality of assessment and grading practices, and (2) why professional training has had little impact. Discrepancies between actual practices and "best" practices in grading were noted in 76 percent of the 34 grading issues. A considerable number of these discrepancies indicated that "a lack of teacher training in grading is a possible cause" (Stiggins, Frisbie, & Griswold, 1989). They concluded that in order to understand the problems that seem to characterize grading practices, the first step is that "we begin with much needed research and translate the results into much needed training" (p.14).

In a related study to determine teacher training priorities, and to compare those findings with results of a decade-long study, Stiggins and Conklin (1988) indicated that discrepancies between the actual assessment needs of teachers and the needs addressed or not addressed by teacher training programs suggest that there are three dimensions that need to be reexamined: policies, certification requirements, and teacher training curricula. According to Hills (1991), having a teacher or an administrator in school who has at least basic training in educational measurement, can make a difference, because "one of the most serious problems of evaluation is the fact that a primary means of assessment, the test itself, is often severely flawed or misused" (Hills, 1991, p.541). What further complicates the problem is that teachers' judgment-based assessment lacks clear and explicit criteria due to lack of training, which leads to "vague criteria communicated to students... and poor performance-rating procedures

contribute undependable assessment and inappropriate instructional decisions" (Stiggins, 1988, p.364).

More recently, Stiggins (1991) studied the degree of mismatch between teachers' assessment training needs and the content of actual training. Stiggins analyzed the assessment training curriculum in 27 undergraduate and graduate programs that provide 75 percent of all the teacher training. He found that only 13 (48 percent) of these programs offer any assessment training at all, and only 6 of these 13 stipulate teacher participation in that training as a requirement for graduation. Stiggins continues that "I believe... that over the years—indeed, over the decades—our training has been painfully and chronically misfocused" (p.8).

If the research on teacher training in educational measurement is important, it is equally imperative to investigate the relevance of measurement instruction to teachers' needs. Some of the emerging research on this issue (Airasian, 1991; Gullickson & Hopkins, 1987) highlights some of the limitations of measurement instruction and training for teachers. Airasian (1991, p. 13) states that "the content of most measurement textbooks and courses shares little relevance with the day to day life of teachers in classrooms. Such a disparity might have been acceptable twenty years ago but it is not today". To support his thesis Airasian examined the current status of educational measurement instruction through the topics covered in educational measurement courses and textbooks and arrived at three main conclusions:

(1) the topics more commonly covered are, for example, the role of educational measurement, educational objectives, item writing, validity, reliability, and checklists and rating scales;

(2) course and textbook discussion of these topics concentrates on formal types of classroom measurement, and

(3) few realistic examples of the classroom use of measurement techniques characterize the vast majority of textbooks. The concepts are presented sometimes in an abstract manner, which makes it unclear how to make these few real-life examples operate in the classroom environment.

Gullickson and Hopkins (1987) surveyed professors from 33 colleges in South Dakota regarding the nature of educational measurement instruction provided students in preservice teacher education programs. Their results may serve as a basis for suggestions concerning the limitations of educational measurement instruction offered to teachers. Their findings indicated that 50 percent of students take a measurement and evaluation course, and the other fifty percent receive it as part of another course. Although these measurement courses are typically taught by professors (81 percent of them hold doctoral degrees), most of them, according to Gullickson and Hopkins, have not majored in educational measurement. These results reinforce "the need for specific understanding of measurement as it applies to classroom practices" (Gullickson & Hopkins, 1987, p.15). Boothroyd, McMorris, and Pruzek (1992, April) found tha "permanently certified teachers were more likely to have completed measurement courses than were provisionally certified teachers", and that "most of those who had completed measurement courses estimated that only a small proportion of the course content directly assisted them in test construction" (p.4).

At the University of Bahrain, all undergraduate students in the college of education are required to take a course in educational measurement and evaluation. In addition, high diploma (one year of full-time study beyond the bachelors degree) and masters degree students in

education are required to take a higher level course in educational measurement and evaluation (University of Bahrain, 1991 -1992).

Obviously, the problem of adequate and relevant preparation and training for teachers in educational measurement is not confined to the United States, but extends to other countries. As a case in point, while extensive use of tests heavily dominates and characterizes Japanese schools, research on testing in Japan seems to be scant (Yamagishi, 1991). Yamagishi highlights the problem as she states that

> Japanese teachers, in general, are not well trained and prepared for tests and evaluation in classrooms, since a specialized course in measurement and evaluation is not required of students enrolled in teacher training programs. Beginning teachers' knowledge of testing is limited to what is provided by a brief review section of a general educational psychology course (p. 170).

Results from a survey conducted by the Japan Educational Psychology Association revealed, according to Yamagishi, that only 28.3% (n was not reported) of the surveyed teachers reported that their professor covered testing and evaluation in classroom. Further, Yamagishi continues that "theoretical and technical issues relevant to classroom evaluation are generally not covered. This suggests that Japanese teachers are generally poorly prepared in testing and evaluation" (p.174).

Teachers' Competency in Educational Measurement

Ebel (1961, p.67) stated that

> It is of the utmost importance to educational progress that the competence of teachers to measure educational attainment be improved.... I am fully persuaded that the current problem in testing which most urgently requires the attention of all professional educators is that of improving the tests made and used by the classroom teacher. To establish the importance of improving teacher competence in measurement, it is necessary to show not only that measurement of educational attainment is needed but also that teachers are currently deficient in getting this job done.

It seems that Ebel's words are still relevant since recent research (e.g., Stiggins, 1991) calls for similar ideas and raises similar concerns.

Generally speaking, research on teachers' competency in educational measurement has taken two dimensions. The first depicts those studies that consist of indirect assessment methods represented by surveying teachers and having them provide self-report ratings of their competency in educational measurement. The second involves more direct assessment of teachers' competency in educational measurement.

Ebel (1961) noted seven errors teachers frequently make in measuring educational achievement. First, teachers rely heavily on their own subjective standards in judging educational achievement although few teachers collaborate in test construction. Second, teachers tend to leave

classroom test preparation until the last minute. Third, teacher-made tests are too short, poorly planned, and lack adequate sampling of the content domain. A fourth error is that teachers tend to place too much emphasis on unimportant details in their classroom tests, while neglecting fundamental principles. Ambiguous teacher- made test items, and rarely having these tests reviewed by a competent colleague is the fifth error Ebel found in teacher-made tests. Sixth, many teachers usually underestimate the effects of the sampling errors on test scores, thereby assuming that the individual student's test score is absolutely accurate. Finally, Ebel reported that teachers, in assessing effectiveness of their own tests, seldom compute statistics on their classroom tests, such as the mean, standard deviation, or the reliability of test scores. Based on the problems he identified in teachers' construction of their classroom tests, Ebel (1962) developed a list of ten measurement principles as a guide for teachers in their classroom testing practices. It seems that the most important among these principles is the one that says: "to measure achievement effectively the classroom teacher must be (a) a master of the knowledge or skills to be tested, and (b) a master of the practical arts of testing" (p.22).

Mayo (1970, p.2) found that "measured competency in measurement was mediocre among not only graduating seniors in teacher training but among the same persons two years after graduation". One would speculate, according to Mayo, that this low performance was probably related to a "lack of commitment to some related factors like mathematics and statistics. He concluded that"... professors should search for teaching innovations in measurement courses" (p.2).

Among the first comprehensive studies designed to assess teachers' competency in educational measurement was Mayo's (1967)

study on preservice preparation of teachers in educational measurement. Mayo developed and administered a measurement competency test to a sample of 2,877 senior education students at 86 institutions. The examinees' average scores (M=28.61, SD=7.284, form A; M=24.97, SD=6.226, form B, Total score=50) led Mayo to conclude that beginning teachers in general do not possess or demonstrate a high degree of knowledge and skills in educational measurement. Mayo concluded that "two possible obstacles impeding improvement of the measurement competency of teachers may be (1) the lack of deep commitment to problems and practices in evaluation, and (2) negative attitude toward statistics" (p.63).

In a related study, Noll (1955) asked 77 college seniors and 108 experienced teachers some questions on fundamental concepts and procedures in measurement. His results showed serious lack of understanding of the basic concepts in measurement among those surveyed.

Newman (1981) found that the teachers surveyed (n=280) averaged 53.7 percent correct responses on her measurement test. She noted that teachers with high scores on the measurement test tended to have better understanding on the affective domain portion than low-scoring teachers. She also indicated that high-scoring teachers appeared to better understand statistical concepts than low-scoring teachers. Newman found that, in general, the teachers surveyed, and low-scoring teachers in particular, view statistical concepts as unimportant.

Ward (1980, 1982), in surveying classroom teachers' knowledge of testing identified three areas in testing in which teachers felt knowledgeable, namely; interpreting test score results, using tests for instructional planning, and judging the relevance of tests.

Using factors such as teachers' knowledge of tests and testing, and the nature and quality of information yielded by tests, Gullickson (1984) surveyed 450 teachers representing different grade levels. He indicated that although teachers perceive themselves as having sufficient knowledge of testing, they are not necessarily knowledgeable in test construction. Gullickson concluded that teachers have limited knowledge in educational measurement. These findings appear to be comparable with other research regarding lack of competence in educational measurement among classroom teachers (Green & Stager, 1986-1987; Newman, 1981; Newman & Stallings, 1982), which suggests that teachers' competency in educational measurement has not changed since Mayo's (1967) study.

It seems reasonable to believe that many preservice and in-service teachers depend on a "repertoire of limited and uninformed test construction skills when they create assessment items" (Carter, 1984, p.57). Carter (1984) interviewed 310 teachers focusing on: (a) teachers' perceptions of their item-writing ability, and (b) teachers' practices in developing items for classroom tests. The study showed that teachers appeared unaware of various item-writing principles presented to them, that their classroom tests were loaded with item-writing problems, and that these principles were not covered as part of their preservice training in educational measurement. The teachers reported that they typically paraphrased textbook materials, and rarely revised their test items. Moreover, Carter noted that teachers incorrectly matched 20 percent of the items to their respective objectives. He also found that the mismatch between items and objectives increased as the level of the objective increased.

More recent research on teachers' needs and proficiencies in testing and measurement (Marso & Pigge, 1987, 1989, 1991; Pigge &

Marso, 1988) has shown some results that may confirm previous research on teachers' competency in educational measurement. For example, Marso and Pigge (1989) surveyed teachers, principals, and supervisors concerning classroom teachers' testing proficiencies. On average, they found no significant difference among teachers in their testing proficiencies as perceived by teachers themselves with regard to grade level and years of teacher experience. Marso and Pigge concluded that more attention should be paid to test item-writing skills that function at cognitive levels beyond the simple knowledge levels in teacher training programs. It can be seen from these results that teachers' proficiencies in test construction are inadequate to meet their needs, and that although principals and supervisors agreed on teachers' evaluation competency needs, they disagreed on teachers' proficiency in educational measurement (Marso & Pigge, 1987). Further, a number of errors were found in item format and construction (Pigge & Marso, 1988), as well as guideline violations which suggests that teachers' test construction skills are weak, inadequate and need improvement to function at higher cognitive levels (Marso & Pigge, 1991).

Educational Measurement Needs of Teachers, Prospective Teachers, Curriculum Specialists, and Other Supervisory Personnel

Both early and recent research have used a variety of approaches in an attempt to identify the essential competencies that teachers, prospective teachers, curriculum specialists and other supervisory personnel need in educational measurement.

Mayo (1964) developed what he called "an ideal list" of 70 competencies in educational measurement. He then surveyed a sample of

"experts" (teachers, principals, superintendents, college and university professors, and testing and research specialists) to see what they thought beginning teachers should know about educational measurement (by rating these competencies in terms of their importance for teachers). He found that, in general, the vast majority of these competencies were considered as important by these "experts". Among the competencies rated as highly important for beginning teachers were knowledge of general principles of test construction, knowledge of the advantages and disadvantages of teacher-made tests, knowledge of effective marking procedures, ability to interpret achievement test scores, knowledge of test administration techniques, ability to state measurable educational objectives, and ability to interpret diagnostic test results to evaluate student progress. It is surprising to know that some essential topics for teachers, such as reliability, validity, item analysis, and statistics, were rated as moderately important by the "experts" surveyed.

In a related but more recent study on teacher-perceived needs in educational measurement and evaluation, Gullickson (1986) noted that teachers usually do not use statistical analysis of tests. Despite the fact that preservice measurement instruction has a relatively substantial emphasis on statistics, according to Gullickson, teachers may devalue its importance because it does not result in a level of understanding that would help them apply statistics in their evaluations. This, however, may be due to lack of a good grasp of statistical concepts (Rudman, Kelly, Wanous, Mehrens, Clark, & Porter, 1980).

Ebel (1962) suggested several competencies in educational measurement that teachers should know, such as knowledge and limitations of tests, the criteria to judge test quality and how to get evidence related to

these criteria, test planning and item writing, test administration procedures, and test score interpretation.

Goehring (1973) reviewed 66 textbooks in educational measurement to identify what competencies are focused in undergraduate educational measurement courses. In addition, he conducted a task analysis of the work of classroom teachers with respect to these competencies. Goehring found that about 67 percent of the competencies identified constitute a major portion of the evaluation competencies that are essential for in-service teachers, such as test construction, application and interpretation of tests, and knowledge of statistical procedures. Goehring concluded that, according to the texts surveyed, measurement courses should concentrate on teacher competencies related to test construction, application, and interpretation of classroom tests as well as on those competencies that have direct effect on pupils and parents.

Stiggins and Conklin (1992) stated that in order to better serve teachers in classroom assessment training, we need to develop a special course tailored to the needs and unique demands of classroom assessment. This course, according to Stiggins and Conklin, will serve teacher training needs by providing teachers with skills and assessment tools they need for their daily classroom assessment demands. Stiggins and Conklin continue that we must plan to provide assessment training and technical assistance to those teachers who completed undergraduate and graduate programs that lack assessment training.

Teachers' knowledge of some core issues in educational measurement, such as validity and reliability, needs to be established. It is important for teachers, for example, to understand how to gather evidence for validity, and that validity is a matter of degree. Teachers should also

know about the role and importance of the reliability of test scores (Merwin, 1989). These issues, however, will be incomplete without teachers' knowledge and skills in elementary statistical methods, and teachers' ability to use "educational assessment to solve instructional problems" (Nitko, 1991, p.2). To be effective, teachers may need to be exposed to a two-year educational measurement program beyond their undergraduate program, along with laboratory work and experience to develop and practice the necessary skills (Schafer & Lissitz, 1991; Nitko, 1991).

In planning a relevant and practical educational measurement training program for teachers, some interrelated factors need to be addressed within the context of training and its linkage to schools. As Stiggins (1988) has suggested, such a training program must include priorities for teachers like "higher order reasoning skills, writing quality paper and pencil test items, integrating assessment and instruction in oral questioning strategies, and designing quality performance assessment based on the observation and professional judgment" (p.15). But to facilitate this training for teachers greater knowledge on the part of researchers and teacher training institutions of the nature of the classroom assessment environment and its demands is needed (Stiggins & Bridgeford, 1985). Stiggins (1991) describes the way he approaches teachers' needs in educational measurement in his 10-session training course. Stiggins emphasizes the following: (1) making teachers aware of the meaning and quality of assessment and its importance for students, (2) the importance of designing assessment with clear achievement targets, (3) design and use of paper and pencil assessment tools, (4) assessment of reading proficiency, (5) using observation and professional judgment as classroom assessment tools, (6) writing assessment and its relation to observation and judgment of

achievement-related behaviors, (7) how to develop sound grading practices, (8) norm-referenced standardized achievement tests, and (9) using computer technology in classroom assessment. In addition, Stiggins focuses on related sound assessment factors, such as having a clear target that fits the purposes of classroom assessment. Evaluation of the impacts of his training course in educational measurement on teachers' classroom assessment proficiency and students' learning showed encouraging results. First, he found improvement in student achievement, which manifested itself in "a higher level of student academic success in writing, speaking, higher order thinking and problem solving" (p.1.1). Second, teachers appeared better able to help students get a clear and more complete understanding of the academic achievement target which, according to Stiggins, led to better instruction, and increased student success rate. Third, there was an increase in students' enthusiasm, and a reduction in assessment anxiety "due to clear expectations, more focus preparation for tests, and improved communication with parents and others about achievement targets" (p. 11).

Another element that must be considered in educational measurement programs is training administrators in classroom assessment (Stiggins, 1986, 1988). For example, principals should master the knowledge and skills for high quality classroom assessment. Stiggins feels that principals and other administrators need this measurement expertise to help teachers with their classroom assessment and daily measurement of student growth. Hence, "it would be useful to expand the dialogue about teacher education in measurement to include the specific assessment skills needed by those in supervisory roles" (Schafer, 1991, p.6), such as principals and curriculum specialists who offer instructional leadership in schools. Thus, teachers' knowledge of educational measurement should be enhanced by

those who communicate with teachers and influence their instructional environment.

Our attitude toward relevant, integrated, and quality measurement programs for teachers should not preclude us from emphasizing the fact that **"the adequate measurement of achievement requires not only skills and scholarship pertaining to test theory and interpretation but also a** sophisticated grasp of the academic content to be measured and an intimate knowledge of appropriate classroom activities" (Rudman, 1987b, p. 10).

Most educators who are responsible for teacher training and education "function in a world that is formal, precise, well articulated, prescriptive, technical, content oriented, and conceptually dominated" (Airasian, 1991, p. 14). Teachers, however, deal with real problems in their daily classroom instruction and assessment. Thus, it is insufficient to identify teachers' needs in educational measurement, but also, as Airasian reports, "we need workshops that are intended to identify what measurement specialists need to know about life in classrooms" (p.14).

In addition to the relevant topics in educational measurement previously discussed, Airasian (1990,1991) calls for other measurement issues and activities that he deems important for teachers, and that need more attention in texts and instruction. For instance, Airasian states that when, at the beginning of the school-year, teachers get to know their students with different abilities, interests, and school experiences, they need to gather some information based on their observation, integrate this information with their expectations and attitudes, and make judgments about the characteristics of individual students. Teachers, then, arrive at what Airasian calls “sizing up” judgments. These judgments need informal observation. Therefore measurement textbooks should address this kind of

observation methods, and should show teachers how the information they collect based on this kind of observation can be incomplete, biased, unstable, suffer from logical errors, and not valid for making appropriate inferences about student performance. Moreover, there are three more measurement activities that Airasian suggests that should be addressed in measurement courses. They are:

(1) Much time should be spent on how teachers can use measurement information to plan instruction and critique instructional materials;

(2) Informal performance assessment should be done regarding how professors' instruction in measurement is going and how teachers feel toward course and professor; and

(3) The packages that accompany school textbooks, such as worksheets, home activities, and achievement tests should be discussed in measurement courses.

"Measurement books should...provide information to help teachers make decisions about the suitability, appropriateness, and validity of such materials for their own instruction" (Airasian, 1991, p.15), because these activities and materials consume more of a teacher's time and may have a more profound impact on learning and instruction than that of the formal measurement procedures emphasized in measurement textbooks. Hence, classroom assessment courses should address these assessment procedures and show the appropriate tools to be used to carry out such classroom assessment tasks. Airasian draws our attention to some other relevant topics which should be addressed in measurement courses for teachers, such as how to use information to plan instruction, how to assess learning during instruction, and how to evaluate instructional materials.

The development of the **Standards for Teacher Competence in Educational Assessment of Students** by the American Federation of Teachers (AFT), the National Council on Measurement in Education (NCME), and the National Education Association (NEA) (1991) has essentially stemmed from the growing concern about the inadequacy of teacher preparation in assessment.

These standards (pp.30-32) are as follows:

1. Teachers should be skilled in choosing assessment methods appropriate for instructional decisions.
2. Teachers should be skilled in developing assessment methods appropriate for instructional decisions.
3. Teachers should be skilled in administering, scoring and interpreting the results of both externally-produced and teacher-produced assessment methods.
4. Teachers should be skilled in using assessment results when making decisions about individual students, planning teaching, developing curriculum, and school improvement.
5. Teachers should be skilled in developing valid pupil grading procedures which use pupil assessment.
6. Teachers should be skilled in communicating assessment results to students, parents, other lay audiences, and other educators.
7. Teachers should be skilled in recognizing unethical, illegal, and otherwise inappropriate assessment methods and uses of assessment information.

These standards seem important, relevant, and integrated with teachers' needs in classroom assessment. I think that these standards, when worded in their final form based on practice and experience, should be addressed in

measurement courses and communicated to teachers both theoretically and practically.

Conclusions

The following conclusions are drawn from the review of literature discussed above:

1. There is an inextricable relationship between testing and teaching, and this relationship should be emphasized in educational measurement courses in teacher-training programs.
2. Teachers spend up to one third of their professional time on testing-related classroom activities.
3. There are many errors in teacher-made tests.
4. There is a discrepancy between teachers' perceived proficiency and their actual competency in educational measurement.
5. Teachers' competency in educational measurement has not changed since Mayo's (1967) study.
6. Educational measurement textbooks do not explicitly emphasize how teachers can integrate testing with instruction.
7. Most research has concentrated on teachers' needs as they perceive them or as they are perceived by principals and supervisors in educational measurement, rather than from directly measuring teachers' competency in educational measurement.
8. The available review of literature on teachers' competency in educational measurement does not appear to have addressed curriculum specialists' needs in this important area.
9. Recent research calls for preparation and assessment of principals and curriculum specialists in educational measurement.

10. Recent research calls for an urgent need to make those responsible for teacher education programs more aware and more knowledgeable about the classroom environment in order to familiarize themselves with the daily classroom problems, and to help them plan their programs on a reality-check basis.

11. Recent research suggests that some educational measurement courses should emphasize informal classroom assessment.

12. Practical educational measurement training for prospective teachers is infrequent. This also applies to those in supervisory positions, such as curriculum specialists.

Descriptive statistics that relate directly to classroom teachers' daily work should be emphasized. Hence, relevance of measurement and statistical topics should be important in order to gain teachers' attention. Teachers' avoidance of measurement courses may be attributed not only to teachers' "dislike" of statistics or to difficulty of measurement courses, but also because teachers deem these courses irrelevant to their daily instruction and assessment of their students (Farr & Griffin, 1973). Teachers' poor comprehension of statistics (Rudman et al., 1980) may be another factor that leads teachers to dislike statistics and make them consider statistics irrelevant to their daily instruction and assessment. Finally, measurement courses should be more practical. Teachers should receive training on how to apply classroom assessment and use it as a vehicle for making sound decisions. Further, teachers need to know how to use and integrate cognitive theories of learning with classroom assessment. This will be a remote target to achieve until measurement courses address it both theoretically and practically. Practical workshops in educational measurement for prospective teachers appear to be scant. If we believe, for

example, that reliability, validity, and item analysis are essential for teachers, then we should help them practice and apply these principles. Moreover, we should show them how to interpret the results of applying these principles in different learning situations, for different purposes, and to make relevant, accurate, correct, and desirable inferences that help teachers make sound decisions about learning and instruction.

CHAPTER III

RESEARCH DESIGN AND PROCEDURE

Introduction

This study regarding the competency of Bahrain teachers, curriculum specialists, and prospective teachers in educational measurement investigated the level of competency in educational measurement, and the strengths and weakness of each group in specific areas of educational measurement (e.g., item analysis, reliability, validity). This chapter addresses the research instruments used, the sampling procedures, the data collection process, the follow-up of non-respondents, data analysis, and the characteristics of the subjects.

Research Instruments

Two research instruments were used in this study: an educational measurement test (hereafter referred to as the "TEST") and a questionnaire. The TEST consisted of 65 multiple-choice items. A total of 50 test items were taken from Boothroyd (1990). Very slight modifications were made on some of those items in order to make the context of the items familiar to the subjects of the study. For example, if the words "New York " appeared in the stem of an item, they would be replaced by the word "Manama" (the capital city of Bahrain). Otherwise, no other modifications were made. The remaining 15 test items were constructed based on measurement and test

construction criteria addressed in measurement textbooks, such as Mehrens and Lehmann (1991) and Wiersma and Jurs (1990).

To maintain the content-related validity of the TEST items, and to permit valid inferences from the TEST, the researcher asked professors of educational measurement at the University of Bahrain to respond to a table of specifications by writing the percentage of instructional time they devoted to each of the 13 measurement areas in their measurement course.

Based on the responses of the measurement professors at the University of Bahrain, the researcher deleted some items and added some other items to the TEST to insure the content-related validity of the measurement test according to the content domain of the measurement courses taught at the University of Bahrain. The TEST was then given to a doctoral candidate in measurement, evaluation and research design at Michigan State University to check for any flaws that might impede or negatively affect any inferences drawn from the test. In addition, the researcher gave the English version and the Arabic version of the TEST to four doctoral students at Michigan State University to check for any errors in translation of the test items from English to Arabic that would change the meaning and result in incorrect inferences. Finally, the researcher gave both the Arabic and English versions of the test to the Department of Arabic at Michigan State University to verify the Arabic translation of the English version of the test. Based on the recommendations of the five doctoral students and the Department of Arabic, the researcher made some very slight changes to the wording of some of the TEST items.

The questionnaire consisted of 23 questions. The selection of the questions was based mainly on prior research findings (Boothroyd, 1990;

Newman, 1981; Mayo, 1967) which showed that these variables are valid and related, in one way or another, to competency in educational measurement. In addition to the demographic data, the questionnaire included items asking for information on, for example, classroom testing practices, item types used in classroom tests, the amount of time devoted to testing, item analysis, current competency in educational measurement, and practical training needed in educational measurement.

Subjects and Sampling Procedures

The target population consists of in-service teachers, curriculum specialists, and prospective teachers in Bahrain. The in-service teachers (n=1300) were randomly sampled by strata (primary, intermediate, and secondary school). Because of their small number, the researcher studied the total population of curriculum specialists (N=146) and prospective teachers (N=221).

Table 3.1 shows the sample of teachers classified by sex and grade level. The percentage of female teachers exceeds that of male teachers both by individual grade level and across grade levels. The largest sex difference, as one might expect, favored the percentage of females at the primary level.

Table 3.1: Information on the sample of teachers by sex and grade level.

Grade level	Male	Female	Total
Primary	334 (44%)	420(56%)	754(58%)
Intermediate	76(49%)	80(51%)	156(12%)
Secondary	190(49%)	200(51%)	390(30%)
Total	600(46%)	700(54%)	1300

Table 3.2 shows the number of participating schools for the sample of teachers by grade level and sex. Table 3.3 shows the total population of curriculum specialists and prospective teachers who participated in the study.

Table 3.2 : Information on the participating schools by sex and grade level.

Grade level	Male	Female	Total
Primary	33(40%)	49(60%)	82(74%)
Intermediate	6(46%)	7(54%)	13(12%)
Secondary	6(38%)	10(62%)	16(14%)
Total	45(41%)	66(59%)	111

Table 3.3 : Information on the total population of curriculum specialists and prospective teachers used in the study.

Group type	Male	Female	Total
Curriculum Specialists	95(65%)	51(35%)	146
Prospective Teachers	10(5%)	211(95%)	221

Characteristics of the subjects

The two research instruments were administered to 1300 teachers, 146 curriculum specialists, and 221 prospective teachers. Table 3.4 below shows the response rate for the total sample and for each group. The highest (82%) response rate was for the teachers who constitute 83.4 percent of the total sample.

The distributions by sex and age are given in Tables 3.5 and 3.6. The mean age of the respondents in the total sample was 34.3 years with a standard deviation of 8.1 years. The highest mean age among the three

Table 3.4: Information on the response rate for the total sample and for each group.

Sample	Responses
Teachers	1070(82%)
Curriculum Specialists	98(67%)
Prospective Teachers	115(52%)
Total	1283(77%)

groups was that of the curriculum specialists (42.4 years). With the exception of the curriculum specialists, there were more females than males in the total sample and in each group independently.

Table 3.5: Distribution of the respondents' age in the study.

Group	n	M	SD	SE	MIN	MAX
Teachers	1070	34.9	7.2	.2	21	60
Curriculum Specialists	98	42.4	5.9	.6	31	57
Prospective Teachers	115	22.2	1.5	.1	20	29
Total Sample	1283	34.3	8.1	.2	20	60

Table 3.6: Distribution of the respondents' sex in the study.

Group	Male	Female	No Answer
Teachers	485(45.3%)	583(54.5%)	2(.2%)
Curriculum Specialists	69(70.4%)	26(26.5%)	3(3.1%)
Prospective Teachers	8(7%)	107(93%)	0(0%)
Total Sample	562(43.8%)	716(55.8%)	5(.4%)

The distribution of highest degree held among respondents (Table 3.7) shows the following: First, only 0.1 percent of the teachers hold a doctorate, 1.0 percent of them have a masters degree, and more than half of the teachers hold at least a bachelors degree (56.6 percent). Second, nearly all of the prospective teachers (99.1 percent) are candidates for a bachelors degree. The 25 percent of the teachers who hold a diploma (associate degree) might be those who graduated from the Teacher Training College in Bahrain between the mid-sixties and early-seventies before the establishment of the University of Bahrain in the early seventies. Among the curriculum specialists 25.5 percent have a bachelors degree, 39.8 percent have a masters degree, and 16.3 percent have a doctorate. The year highest degree was obtained ranged from 1958 to 1992 for the teachers, from 1957 to 1992 for the curriculum specialists, and from 1992 to 1993 for the prospective teachers.

Table 3.7: Distribution of the respondents by highest degree held.

Degree	Teachers	Curriculum Specialists	Prospective Teachers
Doctorate	1(.1%)	16(16.3%)	0(0%)
Masters	11(1%)	39(39.8%)	0(0%)
High Diploma	107(10%)	17(17.4%)	0(0%)
Bachelors	606(56.6%)	25(25.5%)	114(99.1%)
Diploma	267(25%)	1(1%)	0(0%)
High School Certificate	70(6.5%)	0(0%)	0(0%)
Less Than High School Certificate	5(.5%)	0(0%)	0(0%)
No Answer	3(.3%)	0(0%)	1(.9%)

Most recent year when the highest degree held was received was 1986 (6.3 percent) for the teachers, 1986 (9.2 percent) for the curriculum specialists, and 1992 (23.5 percent) for the prospective teachers. Most of the respondents in each group stated that their highest degree held was obtained from a college of education (see Table 3.8).

About 34 percent of the curriculum specialists and 37.1 percent of the teachers stated that education was their major area of study when in the bachelors degree program (see table 3.9).

Table 3.8: Information on the highest degree held obtained from a college of education.

Group	Yes	No	No Answer
Teachers	753(70.4%)	240(22.4%)	77(7.2%)
Curriculum Specialists	82(83.7%)	14(14.3%)	2(2%)
Prospective Teachers	91(79.1%)	0(0%)	24(20.9)

Table 3.9: Distribution of the respondents whose major was education in the bachelors program.

Group	Yes		No		No Answer	
	f	%	f	%	f	%
Teachers	397	37.1	356	33.3	317	29.6
Curriculum Specialists	33	33.7	60	61.2	5	5.1
Prospective Teachers	79	68.7	0	0%	36	31.3

The distribution of the respondents' college major is shown in Table 3.10. The largest percentage among the major subjects across the three

Table 3.10: Distribution of the respondents by major in college.

Major	Teachers	Curriculum Specialists	Prospective Teachers
Islamic Education	15(1.4%)	4(4.1%)	--
Arabic	268(25%)	24(24.5%)	--
English	119(11.1%)	3(3.1%)	--
Mathematics	123(11.5%)	11(11.2%)	--
Science	45(4.2%)	--	--
Chemistry	24(2.2%)	2(2%)	--
Physics	14(1.3%)	3(3.1%)	--
Biology	20(1.9%)	4(4.1%)	--
Engineering	37(3.5%)	7(7.1%)	--
Accounting	11(1%)	2(2%)	--
Business Adm-inistration	12(1.1%)	1(1%)	--
Elementary Teacher	121(11.3%)	3(3.1%)	103(89.6%)
Physical Education	37(3.5%)	4(4.1%)	--
Fine Arts	24(2.2%)	7(7.1%)	--
Music	8(.7%)	1(1%)	--
Psychology	17(1.6%)	4(4.1%)	--
Geography	21(2%)	3(3.1%)	--
History	27(2.5%)	3(3.1%)	--
Economy	9(.8%)	1(1%)	--
Other	79(7.4%)	11(11.2%)	--

groups is Arabic followed by mathematics. The results indicate that 11.3 percent of the teachers major in "class teacher" (elementary teacher), while most of the prospective teachers (in reality all of them) specialize in the "class teacher" major. The next highest percentages are for the major subject labeled "other": 7.4 percent, and 11.2 percent for the teachers, and curriculum specialists, respectively. The "other" major subjects include areas such as French, Library Science, Sociology, Home Economics, and Educational Resources.

Only .6 percent of the teachers majored in Computer (not shown in Table 3.10). Less than 4 percent of the teachers (3.2%) , and 10.4 percent of the prospective teachers did not provide an answer about their major in college.

The distribution for each group regarding their teaching/ supervision subject is displayed in Table 3.11. As might be expected, the results show that the percentages in tables 3.10 and 3.11 are quite similar, especially for Arabic, Mathematics, English, Class Teacher, and Fine Arts.

The results show that only .5 percent of the teachers teach computers (not shown in Table 3.11). 6 percent of teachers, 1.1 percent of curriculum specialists, and 13.9 percent of prospective teachers did not provide an answer.

The distribution of each group by grade level is shown in Table 3.12. The table shows that the highest percentage is in the primary grade level.

Table 3.11: Distribution of the respondents by teaching/ supervision subject.

Subject	Teachers	Curriculum Specialists	Prospective Teachers
Islamic Education	19(1.8%)	4(4.1%)	--
Arabic	268(25%)	23(23.5%)	--
Mathematics	146(13.6%)	12(12.2%)	--
Science	59(5.1%)	6(6.1%)	--
Social Subjects	29(2.7%)	8(8.2%)	--
Elementary Teacher	129(12.1%)	9(9.2%)	99(86.1%)
Geography	11(1%)	--	--
History	11(1%)	--	--
Physics	8(.7%)	1(1%)	--
Chemistry	11(1%)	2(2%)	--
Biology	9(.8%)	--	--
Commerce	27(2.5%)	--	--
Industrial Subjects	42(3.9%)	11(11.2%)	--
English	126(11.8%)	3(3.1%)	--
Fine Arts	27(2.5%)	7(7.1%)	--
Music	8(.7%)	1(1%)	--
Physical Education	37(3.5%)	4(4.1%)	--
Psychology	1(.1%)	--	--
Economy	6(.6%)	1(1%)	--
Other	33(3.1%)	5(5.1%)	--

Table 3.12: Distribution of the respondents by grade level.

Grade	Teachers		Curriculum Specialists		Prospective Teachers	
	f	%	f	%	f	%
Primary	586	54.8	45	45.9	99	86.1
Inter-mediate	145	13.6	19	19.4	--	--
Secondary	298	27.8	24	24.5	--	--
No Answer	41	3.8	10	10.2	16	13.9

Information on the average number of years in teaching for each group is given in Table 3.13. On average teachers have more experience in teaching than curriculum specialists. As expected, teachers have up to 42 years of experience in teaching, while some curriculum specialists have a maximum of 25 years of experience in teaching. In addition, the variability in teaching experience among teachers is slightly higher than that among the

Table 3.13: Distribution of the respondents by the average number of years in teaching.

Group	Mean	Median	SD	SE	MIN	MAX
Teachers	13.1	12.5	7.9	.2	.00	42
Curriculum Specialists	11.9	11	6.1	.7	1	25
Prospective Teachers	.01	.00	.1	.01	.00	1

curriculum specialists. Apparently, the distribution of the range of experience in curriculum supervision among curriculum specialists is close to normal (Mean=9.3, Mode=10, Median=10, with a standard deviation of 4.5 years). The range in years of experience in curriculum supervision for this group is 1- 20 years.

Data Collection

After having the Arabic translation of the English version of the research instruments verified by the Department of Arabic at Michigan Sate University, obtaining permission to conduct the study from the University Committee on Research Involving Human Subjects (UCRIHS) at Michigan State University, and from the officials in the Bahrain Ministry of Education and the University of Bahrain, the researcher traveled to Bahrain to collect the data.

The curriculum specialists were notified about the study and it took them a few hours to complete the materials. The data collection process for teachers, however, was somewhat different. Due to the large sample of teachers (n=1300), and in order to save time and effort, the researcher (with the assistance of a team of 15 curriculum specialists) collected the data. The researcher met with the team for 3 hours, explained to the them the process of data collection in full detail and their role in that process, and assigned them the participating schools. The team of curriculum specialists then personally delivered to their schools the sealed envelopes that contained the appropriate number of booklets of the research instruments, along with a cover letter explaining the goals of the study and requesting the teachers' responses. The school principals were asked to supervise their respective teachers responding to the research instruments, collect the

booklets, and have them ready for the curriculum specialists to collect the following morning. For the prospective teachers, the researcher obtained a list of their names and their respective advisers from the office of the registrar at the University of Bahrain. The researcher contacted the advisers personally and sought their cooperation to have their students respond to the research instruments and return the materials the following day.

Follow-Up of Non-Respondents

As the response rate for the prospective teachers was low compared to that of teachers and curriculum specialists, and to determine whether the results of this study would be affected by non response bias (for the prospective teachers), the researcher attempted to draw a random sample of non-respondents. The University of Bahrain, however, denied the researcher any access to the addresses of non-respondents. In addition, the respective advisers were unable to help in that process. Due to these problems, the researcher is unable to generalize to the total population of prospective teachers.

Data Analysis

The data were analyzed using the SPSSPC. Statistical analyses were conducted to answer the research questions and test the research hypotheses proposed in chapter one. Research questions 1, 2, 3, and 4 (see Chapter 1) were answered by calculating descriptive statistics (e.g., frequencies, mean, standard deviation, standard error of the mean, and percentages). The point biserial correlation coefficient ($r_{pbis)}$ and item difficulty indices (p-values) were computed to supply information on the

performance of respondents on the TEST items. Pearson product moment correlation coefficients were computed to provide information for research hypotheses 1 and 2. Research hypotheses 3 through 12 were tested by conducting a t-test, and one-way and two-way analysis of variance (ANOVA). The KR20 reliability for the TEST scores was computed for the total sample and for each group separately. Multiple regression (MR) analysis was conducted to determine what variables best predict the total score on the TEST. In addition, descriptive statistics (e.g., mean, standard deviation) were also computed. A significance level of .05 was used throughout the study.

Summary

The purpose of this study was two-fold: (1) To assess the current competency of Bahrain teachers, curriculum specialists, and prospective teachers in educational measurement, and (2) to examine their strengths and weaknesses in specific areas in educational measurement. Exploring the factors that affect the respondents' competency in educational measurement, the differences among the three groups in educational measurement, the current classroom testing practices, and the needs for practical training in educational measurement are the issues this study was designed to investigate.

The sample of teachers (n=1300), representing 21 percent of the population of Bahrain teachers (N=6175), was distributed among 111 schools. Although all the available curriculum specialists (N=146) and prospective teachers (N=221) were scheduled to participate in the study, only 98 curriculum specialists and 115 prospective teachers participated.

Descriptive and inferential statistics were used to analyze the data.

CHAPTER IV

ANALYSIS AND INTERPRETATION OF THE DATA

Introduction

The study's findings are presented in these sections. First, the results that deal with classroom testing practices, measurement preparation, and measurement training needs (that pertain to research questions 1, 2, and 3) will be reported. Second, subjects' performance on the TEST (that pertains to research question 4) will be discussed for each group separately. Third, the findings pertaining to the research hypotheses are presented. Some findings of previously conducted research studies, such as Boothroyd's (1990) and Newman's (1981) studies, will be reported where appropriate.

Classroom Testing Practices and Measurement Preparation

Research Question 1: What is the actual preparation of Bahrain teachers, curriculum specialists, and prospective teachers in measurement?
Research Question 2: What are the current testing practices of Bahrain teachers, curriculum specialists, and prospective teachers?

Table 4.1 presents the findings of the amount of measurement coursework completed by each of the three groups. From the table it is seen that most of the prospective teachers (80 percent) had, in contrast to 35.6 percent of

the teachers and 26.5 percent of the curriculum specialists, completed one measurement course. More than one third (36.7 percent) of the curriculum specialists had more than one college course in measurement, whereas only 9.3 percent of the teachers and 4.3 percent of the prospective teachers had more than one college course in measurement. Only 16.2 percent of the teachers and 13.3 percent of the curriculum specialists had in-service training in measurement.

The results of the analysis of the statistics coursework completed are shown in Table 4.2. These results reveal that about one third of the teachers have never been exposed to any kind of independent statistics coursework in contrast to 14.3 percent and 8.8 percent of the curriculum specialists and prospective teachers respectively. Regarding "one college course in statistics", the figures in Table 4.2 show that the prospective teachers had more statistics coursework (43.5 percent) than the curriculum specialists (34.7 percent) and the teachers (21.7 percent). The teachers appear to have the lowest percentage among the three groups who received one college course in statistics. The curriculum specialists had the highest percentage among the three groups who had multiple statistics courses (28.6 percent). This may be attributed to the fact that the curriculum specialists were exposed to more academic coursework than the other groups. The teachers had more in-service training in measurement than they had in statistics (see Tables 4.1, 4.2). This may call for more emphasis on statistics either at the college level or as part of in-service training.

Table 4.1: Distribution of the respondents by measurement coursework completed.

Amount of Coursework	Teachers	Curriculum Specialists	Prospective Teachers
None	162(15.1%)	9(9.2%)	4(3.5%)
Part of another course	129(12.1%)	9(9.2%)	1(.9%)
One college course	381(35.6%)	26(26.5%)	92(80%)
More than one college course	100(9.3%)	36(36.7%)	5(4.3%)
In-service Training	173(16.2%)	13(13.3%)	--
No answer	125(11.7%)	5(5.1%)	13(11.3%)

Table 4.2: Distribution of the respondents by statistics coursework completed.

Amount of Coursework	Teachers	Curriculum Specialists	Prospective Teachers
None	349(32.6%)	14(14.3%)	10(8.8%)
Part of another course	206(19.3%)	9(9.2%)	17(14.7%)
One college course	232(21.7%)	34(34.7%)	50(43.5%)
More than one college course	55(5.1%)	28(28.6%)	21(18.3%)
In-service training	59(5.5%)	6(6.1%)	-- --
No answer	169(15.8%)	7(7.1)	17(14.7%)

From Table 4.3 it can be seen that nearly one-half (46.3 percent) of the teachers stated they spend more than 20 percent of their professional time on classroom testing activities, and about one third of them said that they spend between 11 percent and 20 percent of their professional time on classroom testing activities. These percentages seem to be consistent with

prior research findings (e.g., Stiggins & Conklin, 1988; Stiggins, 1991; Schafer, 1991; Green & Stager, 1986-1987; & Gullickson, 1982). The curriculum specialists, however, devote less time to classroom testing activities than teachers but more than the prospective teachers.

Table 4.3: Distribution of the respondents by percentage of time devoted to classroom testing.

Amount of Time	Teachers	Curriculum Specialists	Prospective Teachers
<5%	30(2.8%)	3(3.1%)	6(5.2%)
5%-10%	178(16.6%)	22(22.4)	19(16.5%)
11%-20%	336(31.4%)	27(27.6%)	25(21.7%)
>20%	495(46.3%)	17(17.3%)	28(24.4%)
No answer	31(2.9%)	29(29.6%)	37(32.2%)

Table 4.4 shows how the three groups use teacher-made tests. Using classroom tests exclusively for each purpose ranged from .5 percent to 5.1 percent across the groups.

While diagnosis, achievement, and mastery for learning, respectively, rank the highest among the purposes for classroom testing for teachers in this study, Boothroyd (1990) found that mastery and understanding of the content was the most frequently purpose cited by teachers, followed by remediation and improving instruction, and grading.

Table 4.4: Distribution of the respondents by purposes of classroom test use.

Purpose	Teachers	Curriculum Specialists	Prospective Teachers
Diagnosis (1)	4.4%	2.0%	4.3%
Achievement (2)	3.6	5.1	.9
Mastery of learning (3)	2.9	1.0	3.5
Academic progress (4)	1.1	---	.9
Motivation for learning (5)	1.7	---	---
Grading (6)	.9	---	---
Planning instruction (7)	.5	---	---
1,2,3	2.1	11.2	7.0
1,2,3,4	2.3	4.1	3.5
1,2,3,6	2.0	1.0	5.2
1,2,3,4,5,6,	4.4	3.1	1.7
1 through 7	6.2	---	1.7

Using only one item type for classroom tests seems to be rare among the groups (see Table 4.5). Using a combination of more than one item type in the classroom is the trend for both prospective teachers and teachers (20 percent and 18.6 percent, respectively) who said that they use true-false, matching, completion, short answer, and multiple choice items on their classroom tests.

Table 4.5: Item types used in classroom (percentages).

Item Type	Teachers	Curriculum Specialists	Prospective Teachers
True-False[1]	.8	--	2.6
Matching[2]	.5	1	.9
Completion[3]	.7	--	.9
Short Answer[4]	2.8	4.1	--
Multiple-Choice[5]	2.8	--	--
Essay[6]	1	--	--
1,2,4,5	1.5	3.1	7.0
1,3,4,5	7.0	1.0	5.2
1,2,3,4,5	18.6	4.1	20.0
1,2,3,4,5,6	10.4	12.2	--

The least used item types, in general, are matching and completion.

Information on item types ranked according to their preference for each group is given in Table 4.6 . The table shows that, by and large, essay item type is the most preferred and multiple choice item type is the least preferred by the three groups. Apparently, rankings of the teachers and prospective teachers with respect to true-false, Multiple-Choice, and essay item types are identical, whereas they are completely different from the rankings found by Reynolds and Menard (1980) and Newman (1981). For example, Newman (1981) found that the completion item type was ranked the highest, followed by Multiple-Choice item type.

Table 4.6: Item types ranked according to preference*.

Item Type	Teachers	Curriculum Specialists	Prospective Teachers
True-False	4	3	4
Matching	2	2	5
Multiple-Choice	6	6	6
Completion	3	1	2
Short Answer	5	5	3
Essay	1	4	1

*Note: 1=Most preferred, 6= Least preferred.

Most of the teachers and curriculum specialists stated that they prepare a test plan and conduct item analysis of classroom tests. The percentage ranged from about 30.4 percent to 76.2 percent for the test plan variable, and from 43.5 percent to 84.1 percent for the item analysis variable (see Tables 4.7 and 4.8). The prospective teachers, however, had the lowest percentage among the three groups regarding the test plan preparation (30.4 percent) and item analysis of classroom tests (43.5 percent).

Table 4.7: Distribution of the respondents by test plan preparation.

Group	Yes	No	No answer
Teachers	815(76.2%)	227(21.2%)	28(2.6%)
Curriculum Specialists	73(74.5%)	14(14.3%)	11(11.2%)
Prospective Teachers	35(30.4%)	52(45.2%)	28(24.4%)

Table 4.8: Distribution of the respondents by item analysis practice.

Group	Yes		No		No Answer	
	f	%	f	%	f	%
Teachers	900	84.1	150	14.0	20	1.9
Curriculum Specialists	78	79.6	8	8.2	12	12.2
Prospective Teachers	50	43.5	36	31.3	29	25.2

Data on self-ratings of each group concerning the general competency in educational measurement and competency in test construction are displayed in Tables 4.9, and 4.10, respectively. Less than one-half (42.2 percent), 30.6 percent, and 21.7 percent of the teachers, curriculum specialists, and prospective teachers, respectively, perceive themselves as very competent in measurement in general. Regarding the perceived competency in test construction, Table 4.10 shows that 46.1 percent of the teachers, 35.7 percent of the curriculum specialists, and 23.5 percent of the prospective teachers believe that they are very competent in test construction. Moreover, subjects in each group tended to view themselves, in general, as more competent in test construction than in educational measurement as a whole.

Table 4.9: Distribution of the respondents' self-ratings regarding their general competency in educational measurement.

Level	Teachers	Curriculum Specialists	Prospective Teachers
Excellent	175(16.4%)	12(12.2%)	1(.9%)
Very good	451(42.2%)	30(30.6%)	25(21.7%)
Good	333(31.1%)	31(31.6%)	46(40%)
Adequate	73(6.8%)	14(14.3%)	14(12.2%)
Poor	11(1%)	1(1.1%)	3(2.6%)
No answer	27(2.5%)	10(10.2%)	26(22.6%)

Table 4.10: Distribution of the respondents' self-ratings concerning their competency in test construction.

Level	Teachers		Curriculum Specialists		Prospective Teachers	
	f	%	f	%	f	%
Excellent	258	24.1	24	24.5	5	4.3
Very good	493	46.1	35	35.7	27	23.5
Good	230	21.5	25	25.5	49	42.6
Adequate	55	5.1	4	4.1	7	6.1
Poor	4	.4	--	--	1	.9
No Answer	30	2.8	10	10.2	26	22.6

Research Question 3: In what areas of educational measurement do curriculum specialists, teachers, and prospective teachers need practical training?
Table 4.11 gives a general indication of the practical training needs

Table 4.11: Distribution of the respondents' practical training needs in educational measurement (percentages).

Area	Teachers	Curriculum Specialists	Prospective Teachers
Test planning	6.0	2.0	1.7
Item construction	1.6	--	.9
Item type selection	3.1	--	--
Item analysis	6.3	7.1	3.5
Validity	1.9	1.0	.9
Reliability	4.6	1.0	--
Score Interpretation	1.7	5.1	3.5
Observation methods	2.2	3.1	--
All areas	1.3	3.1	2.6

in educational measurement. The percentages in the table are small due to low response rate for this question, and may not highlight the needs of the majority in each group.

Performance of the Subjects on the TEST

The KR20 reliability and standard error of measurement (SEM) estimates for the sets of scores from the TEST are given in Table 4.12. The TEST appears to produce reliable scores for the total sample and for each group independently.

Boothroyd (1990) set a standard of .54. Since this study used most of his items, our standard in this study might also be .54. However, since 15 of the TEST items were new, we selected a standard of .50. Hence, our standard reflects the estimated probability of an individual (e.g., a teacher) with "minimal competency in measurement" correctly answering an item.

Table 4.12: KR20 reliability and SEM estimates for the TEST.

Group	KR20	SEM
Teachers	.73	3.63
Curriculum specialists	.83	3.41
Prospective teachers	.75	3.60
Total sample	.78	3.67

Research Question 4: What is the current level of knowledge and understanding of Bahrain teachers, curriculum specialists, and prospective teachers in measurement?

(a) how competent is each group in measurement?

(b) in which measurement area(s) is each group the most and the least competent?

Table 4.13 provides information on the overall performance of each group on the TEST and their performance on each of the 13 areas that comprise the total test. It can be seen from the table that the teachers' total mean score on the TEST was 26.72 (41 percent), the curriculum specialists' total mean score was 37.49 (58 percent), and the prospective teachers' total mean score was 34.07 (52 percent). The scores on the TEST ranged from 9 to 48 for the teachers, from 19 to 50 for the curriculum specialists, and from 18 to 49 for the prospective teachers. It should be noted that the curriculum specialists outperformed the teachers and the prospective teachers in all 13 areas of the TEST, except for the areas of "reliability" and "types of items". The teachers performed better in the areas of "objectives", and "grading and marking" than they did in all the other areas of the TEST. In general, the curriculum specialists performed well in all areas, except for the areas of "types of tests" and "standard error of measurement." The prospective teachers scored better in the areas of "objectives", "types of items", "item writing", "grading and marking", "correlation" "reliability", and "validity" than they did in the other areas of the TEST. The table shows that the teachers are weakest in more areas than the other groups. The teachers need more training in different areas such as "item analysis", "score interpretation" and "standard error of measurement." The curriculum specialists need more training in the areas of "types of tests" and "standard error of measurement." The prospective teachers need more training in the areas of "test planning", "types of tests", "test construction", "item analysis", "score interpretation", and "standard error of measurement". All the three groups together are weak in the areas of "types of tests" and "standard error of measurement". Plake, Impara, and Fager (1992), however, found that teachers (n=555) demonstrated the highest performance in the areas

Table 4.13: Distribution of the respondents' performance on the TEST by content area.

Content Area	# of Items	Teachers			Curriculum Specialists			Prospecti-ve Teachers		
		M	SD	%	M	SD	%	M	SD	%
Test Planning	2	.72	.69	36	1.15	.69	58	.70	.65	35
Objective-s	9	4.63	1.87	51	6.28	1.36	70	5.82	1.80	65
Types of Tests	6	1.98	1.10	33	2.45	1.11	41	2.22	1.09	37
Types of Items	8	3.90	1.59	49	4.87	1.43	61	5.12	1.50	64
Item Writing	7	2.96	1.50	42	4.40	1.50	63	4.18	1.56	60
Test Construc-tion	4	1.78	.91	45	2.39	.84	60	1.86	1.00	46
Item Analysis	5	1.54	1.06	31	2.54	1.28	51	2.00	1.07	40
Score Int-erpretatio-n	8	2.14	1.28	27	4.12	2.00	52	2.94	1.66	37
Grading & Marking	3	1.85	.92	62	2.31	.74	77	1.67	.81	56
Correlatio-n	3	1.21	.85	40	2.00	.83	67	1.77	1.01	59
Reliability	5	1.88	1.16	38	3.00	1.34	60	3.09	1.17	62
Standard Error	3	.89	.70	30	1.12	.63	37	.94	.72	31
Validity	2	.97	.66	48	1.47	.63	74	1.20	.63	60
Total Test	65	26.72	6.99	41	37.49	8.27	58	34.07	7.19	52

of administering, scoring, and interpreting test results.

Item analysis data for the TEST for each group are displayed in Tables 4.14 through 4.16 (see Appendix F). The TEST item difficulties (p-values), indicating the proportion in each group answering an item correctly, ranged from .10 to .79 with an average difficulty of .40 for the teachers, from .00 to .94 with an average difficulty of .57 for the curriculum specialists, and from .03 to .90 with an average difficulty of .57 for the prospective teachers. Point biserial correlation coefficients, which are an index of item

discrimination, ranged from -.23 to .43 with an average correlation of .16 for the teachers, from -.29 to .58 with an average correlation of .23 for the curriculum specialists, and from -.32 to .64 with an average correlation of .18 for the prospective teachers.

The most discriminating items, those with a discrimination level of .40 or above, were items 9 and 33 for the teachers, items 9, 14, 15, 20, 27, 33, 38, 39, 41, 44, 45, 49, 52, 57, and 63 for the curriculum specialists, and items 7,19, 20, 33, 39, 44, 48, and 62 for the prospective teachers. In general, the lowest performance on the TEST across the three groups was on item 56.

The overall perceived needs of the respondents regarding practical training in educational measurement (see Table 4.11), especially in the areas of item analysis and standard error of measurement, may reflect their weakness in these areas based on their performance on the TEST (see Table 4.13).

Testing the Research Hypotheses

Research Hypothesis 1: For each of Bahrain teachers, curriculum specialists, and prospective teachers there is no significant relationship between their measurement competency and highest degree held, recency of highest degree held, major subject of specialization, major subject of teaching/ supervision, grade level of teaching/ supervision, years of experience in teaching, years of experience in supervision, amount of measurement coursework completed, recency of measurement coursework completed, amount of statistics completed, recency of statistics completed, percentage of professional time devoted to test construction, and purposes of using teacher-made classroom tests.

Research Hypothesis 2: For each of Bahrain teachers, curriculum specialists, and prospective teachers there is no significant relationship between their measurement

competency and developing a test plan prior to writing test items, conducting item analysis on self-constructed classroom tests, self-assessment of the level of competency in measurement in general, and self-assessment of the level of competency in test construction principles.

Product moment correlations were computed to investigate the relationship between competency in educational measurement as measured by the TEST, and measurement preparation variables, classroom testing variables, and subjects' background variables. The results of these analyses are displayed in Table 4.17. We should be very cautious in interpreting these findings for the following reason. Despite the fact that some significant correlations were obtained, the proportion of variance in the TEST scores accounted for by any one of these variables was very small; the largest r^2 was .08. In other words, in some cases it is possible that trivial or negligible relationships were significant because of the large sample size. This might be the case with the group of teachers where there were four significant relationships; the sample size of the teachers is the largest among the three groups. In general, the findings in Table 4.17 indicate that the correlations are small or negligible, and some of them are negative. Although the highest correlation coefficient was -.37 (r^2=.14), it is not significant, and negative as well. These negative correlations signify that high values on the variables in the table are associated with low scores on the TEST, and vice versa. Some variables, such as developing a test plan, conducting item analysis, self-assessment of competency in educational measurement, and purposes of using teacher-made classroom tests (for the teachers group), indicated a significant relationship with the scores on the TEST. The results revealed that those teachers who stated they develop a test plan and conduct item analysis scored low on the TEST.

Table 4.17: Correlations between the scores on the TEST and classroom testing activities variables and measurement preparation variables.

Variable	Teachers	Curriculum Specialists	Prospective Teachers
1	-.08	.01	-.12
2	.14	.24	-.10
3	.13	-.07	.
4	.07	-.37	.
5	-.04	-.09	.
6	-.05	-.14	.
7	----	-.18	----
8	-.06	-.06	-.17
9	-.001	.07	-.02
10	-.02	-.02	.02
11	.04	.15	-.24
12	.08	-.12	.05
13	.21**	.18	-.04
14	-.28**	-.16	-.19
15	-.19*	-.21	-.25
16	.18*	.28	.08
17	.14	.32	-.09

*P<.01; **P<.001; . Coefficient cannot be computed; --- Does not apply to these groups. 1=Highest degree held, 2=Recency of highest degree held, 3=Major subject of specialization, 4=Major subject of teaching/ supervision, 5= Grade level of teaching/ supervision, 6=years of experience in teaching, 7=years of experience in supervision, 8=Amount of measurement coursework completed, 9=recency of measurement coursework completed, 10=Amount of statistics completed, 11=Recency of statistics completed, 12= percentage of professional time devoted to test construction, 13=Purposes of using classroom tests, 14=Developing a test plan, 15=Conducting item analysis, 16=Self-assessment of general competency in measurement, 17=Self-assessment of competency in test construction.

Those teachers who perceived themselves as competent in educational measurement and use classroom tests for a myriad of purposes, scored high on the TEST. To see the strength of association between the scores on the TEST and each of these two variables (measurement competency and purposes of using classroom test) for the group of teachers, a one-way ANOVA was conducted on the TEST scores with each of these two variables (see Tables 4.18, 4.19), in addition to the index of association strength (ω^2) computation.

Table 4.18: ANOVA for the TEST scores with "purposes of classroom test" variable for the teachers.

Source	D.F.	Sum of Squares	Mean Square	F-Ratio	F-Prob.	ω^2
Between Groups	6	471.5559	78.5927	1.5087	.1846	.03
Within Groups	**88**	**4584.1915**	**52.0913**			
Total	**94**	**5055.7474**				

From Table 4.18 we see that there is no significant difference in measurement competency between the teachers who use classroom tests for a myriad of purposes and the teachers who use classroom tests for one or two purposes, where the index of association strength has a negligible value (ω^2=.03). This suggests that the shared variance between the TEST scores and "purposes of classroom test use" variable is very small. Hence, measurement competency does not depend on using classroom tests for several purposes. Despite the fact that there is a significant difference between the teachers who perceived themselves as competent and those teachers who perceived themselves as incompetent in measurement (see Table 4.19), the proportion of the shared variance between the scores on the TEST and the perceived competency in measurement is negligible (ω^2=.02). In addition, the Scheffe test showed that the significant difference was only between the teachers who perceived themselves as "excellent" and the teachers who perceived themselves as "good" in measurement.

Table 4.19: ANOVA for the TEST scores with the "perceived level of general competency in measurement" for the teachers.

Source	D.F.	Sum of Squares	Mean Squares	F-Ratio	F-Prob	ω^2
Between Groups	4	957.8273	239.4568	5.0034	.0006	.02
Within Groups	659	31538.6486	47.8583			
Total	663	32496.4759				

Newman's (1981) results seem to be comparable to the results of this study. For example, Newman found that there was a relationship between the scores on the measurement test and variables such as the amount of measurement coursework (r=.10), recency of measurement coursework (r=.15, p<.05), years of teaching experience (r=-.14, p<.05), highest degree held (r=.13, p<.05), recency of highest degree held (r=.09), percentage of time devoted to classroom testing (r=.06), and purposes of using classroom tests (r=.22, p<.01). The findings of this study showed that the variable "purposes of using classroom tests" has a significant relationship with scores on the TEST (r=.21, p<.001) for the teachers. The conclusion regarding rejecting or not rejecting the research hypotheses 1 and 2 is as follows: For the teachers, research hypothesis 1 is rejected for the "purposes of using classroom tests" variable and not rejected for the rest of the variables in the hypothesis. The research hypothesis 2 is rejected for the variables "developing a test plan", "conducting item analysis", and

"self-assessment of general competency in measurement". The hypothesis is not rejected for the rest of the variables in the hypothesis. Concerning the groups of curriculum specialists and prospective teachers, the hypotheses 1 and 2 are not rejected for all the variables in these hypotheses. By and large, it can be concluded that measurement preparation variables, classroom testing variables, and subjects' background variables do not have that large of an influence on competency in educational measurement.

Research Hypothesis 3: There is no significant difference in measurement competency among Bahrain teachers, curriculum specialists, and prospective teachers.

The analysis of variance (ANOVA) test results are displayed in Table 4.20b. It is evident that there is a significant difference in measurement competency among Bahrain teachers, curriculum specialists, and prospective teachers. The shared variance (ω^2) between the type of respondent and measurement competency was .18. The Scheffe test indicated that the teachers' mean score on the TEST was significantly lower than the mean score of the curriculum specialists. This may signify that the curriculum specialists are more competent in educational measurement than the teachers (see also Table 4.20a). Hence, the hypothesis is rejected.

Table 4.20a: Distribution of the respondents' mean performance on the TEST.

Group	M	SD	MIN	MAX
Teachers	26.72	6.99	9	48
Curriculum Specialists	37.49	8.27	19	50
Prospective Teachers	34.07	7.19	18	49

Table 4.20b: ANOVA for the TEST scores with the "type of respondent".

Source	D.F.	Sum of Squares	Mean Squares	F-Ratio	F-Prob.	ω^2
Between Groups	2	8726.6013	4363.3006	86.5545	.0000	.18
Within Groups	804	40530.4570	50.4110			
Total	806	49257.0583				

Research Hypothesis 4: There is no significant difference in measurement competency (for each group separately) between those who had measurement coursework and those who did not.

Table 4.21b presents the analysis of variance data (for each group) for the scores on the TEST and measurement coursework completed. From the table it can be seen that there is a significant difference between those who had measurement coursework and those who did not regarding the competency in educational measurement (see also Table 4.21a). The Scheffe test showed that for the teachers, those who completed one

measurement course are significantly different from each of the following: those who had no measurement coursework at all, those who had measurement as part of another course, and those who had in-service training in measurement. For the curriculum specialists, those who had measurement coursework as part of another course appear to be significantly different from: those who had no measurement coursework at all, those who had one measurement course, and those who completed more than one measurement course. For the prospective teachers, the significant difference was found between those who had one measurement course, and those who had no measurement coursework at all. It can be concluded that competency in measurement is partly attributable to the amount of measurement coursework or training in measurement. In spite of these findings, the degree of association between measurement coursework and competency in educational measurement as measured by the TEST is not strong. For example, ω^2 shows that the shared proportion of variance between measurement coursework and competency in educational

Table 4.21a: Distribution of the respondents by measurement coursework completed.

Amount of Coursework	Teachers	Curriculum Specialists	Prospective Teachers
None	162(15.1%)	9(9.2%)	4(3.5%)
Part of another course	129(12.1%)	9(9.2%)	1(.9%)
One college course	381(35.6%)	26(26.5%)	92(80%)
More than one college course	100(9.3%)	36(36.7%)	5(4.3%)
In-service Training	173(16.2%)	13(13.3%)	--
No answer	125(11.7%)	5(5.1%)	13(11.3%)

measurement is 7 percent for the teachers, 24 percent for the curriculum specialists, and 9 for the prospective teachers. Therefore, this hypothesis is rejected for each group independently.

Table 4.21b: ANOVA for the TEST scores with "measurement coursework".

TEACHERS

Source	D.F.	Sum of Squares	Mean Squares	F-Ratio	F-Prob.	ω^2
Between Groups	4	2259.8169	564.9542	12.0814	.0000	.07
Within Groups	595	27823.5814	46.7623			
Total	599	30083.3983				

CURRICULUM SPECIALISTS

Source	D.F.	Sum of Squares	Mean Squares	F-Ratio	F-Prob.	ω^2
Between Groups	4	961.7224	240.4306	4.8253	.0026	.24
Within Groups	44	2192.4000	49.8273			
Total	48	3154.1224				

PROSPECTIVE TEACHERS

Source	D.F.	Sum of Squares	Mean Squares	F-Ratio	F-Prob.	ω^2
Between Groups	3	446.0193	148.6731	3.2053	.0287	.09
Within Groups	66	3061.3522	46.3841			
Total	69	3507.3715				

Research Hypothesis 5: There is no significant difference in measurement competency (for each group separately) between those who had coursework in statistics and those who did not.

The data used to test hypothesis 5 are given in Table 4.22b. From Table 4.22b one can conclude that there is a significant difference in measurement competency among both teachers and curriculum specialists depending on the type and amount of coursework in statistics (see also Table 4.22a). The Scheffe test revealed that, for the teachers, those who had no statistics coursework are significantly different from those who had it as part of another course, and those who had one course. For the curriculum specialists, there was a significant difference between those who had no statistics coursework at all and those who had more than one statistics course. The index of association strength (ω^2) shows that the proportion of shared variance between statistics coursework and competency in educational measurement is 4 percent for the teachers, and 16 percent for the curriculum specialists.

Table 4.22a: Distribution of the respondents by statistics coursework completed.

Amount of Coursework	Teachers	Curriculum Specialists	Prospective Teachers
None	349(32.6%)	14(14.3%)	10(8.8%)
Part of another course	206(19.3%)	9(9.2%)	17(14.7%)
One college course	232(21.7%)	34(34.7%)	50(43.5%)
More than one college course	55(5.1%)	28(28.6%)	21(18.3%)
In-service training	59(5.5%)	6(6.1%)	-- --
No answer	169(15.8%)	7(7.1)	17(14.7%)

Table 4.22b: ANOVA for the TEST scores with "statistics coursework".

TEACHERS

Source	D.F.	Sum of Squares	Mean Squares	F-Ratio	F-Prob.	ω^2
Between Groups	**4**	**487.3506**	**371.8376**	**7.8120**	**.0000**	**.04**
Within Groups	569	27083.3550	47.5982			
Total	573	28570.7056				

CURRICULUM SPECIALISTS

Source	D.F.	Sum of Squares	Mean Squares	F-Ratio	F-Prob.	ω^2
Between Groups	4	660.4658	165.1164	3.2409	.0207	.16
Within Groups	43	2190.7842	50.9485			
Total	47	2851.2500				

PROSPECTIVE TEACHERS

Source	D.F.	Sum of Squares	Mean Squares	F-Ratio	F-Prob.
Between Groups	3	295.5489	98.5163	2.0064	.1219
Within Groups	64	3142.5099	49.1017		
Total	67	3438.0588			

No significant difference was found for the group of the prospective teachers. Hence, this hypothesis is rejected for the teachers and curriculum specialists but not rejected for the prospective teachers.

Research Hypothesis 6: There is no significant difference in measurement competency (for each group separately) between those who develop a test plan and those who do not.

From Table 4.23b it is seen that there is a significant difference in measurement competency between teachers who develop a test plan and those who do not. ω^2 was only .06. There is no significant difference in measurement competency between those curriculum specialists and those

Table 4.23b: ANOVA for the TEST scores with "test plan".

		TEACHERS				
Source	D.F.	Sum of Squares	Mean Squares	F-Ratio	F-Prob.	ω^2
Between Groups	1	1929.9029	1929.9029	41.7609	.0000	.06
Within Groups	662	30593.0836	46.2131			
Total	663	32522.9865				
		CURRICULUM SPECIALISTS				
Source	D.F.	Sum of Squares	Mean Squares	F-Ratio	F-Prob.	
Between Groups	1	7.3500	7.3500	.1154	.7356	
Within Groups	46	2928.6500	63.6663			
Total	47	2936.0000				
		PROSPECTIVE TEACHERS				
Source	D.F.	Sum of Squares	Mean Squares	F-Ratio	F-Prob.	
Between Groups	1	23.6686	23.6686	.4204	.5193	
Within Groups	59	3322.0035	56.3051			
Total	60	3345.6721				

prospective teachers who develop a test plan and those who do not (see also Table 4.23a). Therefore, the hypothesis is rejected for the teachers, but not rejected for the curriculum specialists and prospective teachers.

Table 4.23a: Distribution of the respondents by test plan preparation

Group	Yes	No	No answer
Teachers	815(76.2%)	227(21.2%)	28(2.6%)
Curriculum Specialists	73(74.5%)	14(14.3%)	11(11.2%)
Prospective Teachers	35(30.4%)	52(45.2%)	28(24.4%)

Research Hypothesis 7: There is no significant difference in measurement competency (for each group separately) between those who conduct item analysis and those who do not.

The results on Table 4.24b show that there is a significant difference in measurement competency between those teachers who stated they conduct an item analysis and those who stated that they do not conduct item analysis (see also Table 4.24a). Despite this significant difference, the degree of association between the two variables appears to

Table 424a: Distribution of the respondents by item analysis practice.

Group	Yes		No		No Answer	
	f	%	f	%	f	%
Teachers	900	84.1	150	14.0	20	1.9
Curriculum Specialists	78	79.6	8	8.2	12	12.2
Prospective Teachers	50	43.5	36	31.3	29	25.2

be small, where the shared proportion of variance between the two variables (ω^2) was 1 percent for the teachers. The hypothesis therefore is rejected for the teachers, but not rejected for the curriculum specialists and prospective teachers.

Table 4.24b: ANOVA for the TEST scores with "item analysis".

TEACHERS

Source	D.F.	Sum of Squares	Mean Squares	F-Ratio	F-Prob.	ω^2
Between Groups	1	303.5957	303.5957	6.2371	.0128	.01
Within Groups	664	32320.4659	48.6754			
Total	665	32624.0616				

CURRICULUM SPECIALISTS

Source	D.F.	Sum of Squares	Mean Squares	F-Ratio	F-Prob.
Between Groups	1	8.8889	8.8889	.1397	.7103
Within Groups	46	2927.1111	63.6329		
Total	47	2936.0000			

PROSPECTIVE TEACHERS

Source	D.F.	Sum of Squares	Mean Squares	F-Ratio	F-Prob.
Between Groups	1	163.2983	163.2983	2.9762	.0898
Within Groups	58	3182.3517	54.8681		
Total	59	3345.6500			

Research Hypothesis 8: There is no gender by type of respondent (e.g., a teacher) interaction regarding their competency in educational measurement.

Table 4.25 displays the information used to test this research hypothesis. It is evident that there is a significant interaction between respondents' gender and the type of respondent concerning their competency in educational measurement. This may indicate that the conceptual understanding of the subjects in educational measurement from one group to another is affected by the gender of the subject. The hypothesis is rejected.

Table 4.25: ANOVA for the joint effects of gender and type of respondent with the TEST scores.

Source	Sum of Squares	D.F.	Mean Squares	F-Ratio	F-Prob.
Gender	3192.194	1	3192.194	69.812	.000
Type of Respondent	8584.720	2	4292.360	93.872	.000
Interaction	697.966	2	348.983	7.632	.001
Residual	36580.530	800	45.726		
Total	49173.029	805	61.085		

Research Hypothesis 9: There is no type of respondent (e.g., teacher) by respondents' age interaction regarding their competency in educational measurement.

From Table 4.26b it is seen that the F interaction value (between the variables age and type of respondent) is 1.864, $p>.05$. This leads us to

conclude that the competency in educational measurement among teachers, curriculum specialists and prospective teachers does not depend on whether the respondent is old or young (see also Table 4.26a). This hypothesis is not rejected.

Table 4.26a: Distribution of the respondents' age in the study.

Group	M	SD	MIN	MAX
Teachers	34.9	7.2	21	60
Curriculum Specialists	42.4	5.9	31	57
Prospective Teachers	22.2	1.5	20	29

Table 4.26b: ANOVA for the joint effects of age and type of respondent with the TEST scores.

Source	Sum of Squares	D.F.	Mean Squares	F-Ratio	F-Prob.
Age	2521.112	7	360.159	7.477	.000
Type of Respondent	6490.285	2	3245.143	67.371	.000
Interaction	538.791	6	89.799	1.864	.084
Residual	35789.126	743	48.168		
Total	46802.358	758	61.745		

Research Hypothesis 10: There is no significant difference between male and female teachers in their measurement competency.

Table 4.27 indicates that there is a significant difference between male and female Bahrain teachers in their measurement competency ($p<.05$). This may suggest that there is a difference in the conceptual understanding between male and female teachers in different areas in educational measurement. Females may be more competent in educational measurement than males. The hypothesis is rejected.

Table 4.27: t-test for the difference between male and female teachers in educational measurement.

Group	Number of cases	MEAN	SD	SE
Males	303	24.1023	6.372	.366
Females	369	28.8997	6.741	.351

Pooled variance estimates: t-value=-9.41, D.F.=670, 2-Tail Prob.=.000

Research Hypothesis 11: There is no significant difference among teachers at different grade levels in measurement competency.

Table 4.28 shows that there is no significant difference ($p>.05$) among teachers concerning their competency in measurement at the three grade levels. This hypothesis is not rejected. Hence, the grade level has no significant effect on the competency of the teacher in educational measurement.

Table 4.28: ANOVA for the grade level of the teacher with the TEST scores.

Source	D.F.	Sum of Squares	Mean Squares	F-Ratio	F-Prob.
Between Groups	2	107.0321	53.5160	.8893	.4114
Within Groups	**761**	**45795.3802**	**60.1779**		
Total	**763**	**45902.4123**			

Research Hypothesis 12: There is no gender by grade level interaction for the teachers concerning their measurement competency.

ANOVA results for research hypothesis 12 are displayed in Table 4.29b. It appears from Table 4.29b that there is no interaction between the teacher's gender and the grade level at which the teacher practices instruction insofar as measurement competency is concerned. This indicates

Table 4.29b: ANOVA for the joint effects of teacher's gender and grade level with the competency in educational measurement.

Source	Sum of Squares	D.F.	Mean Squares	F-Ratio	F-Prob
Gender	3936.422	1	3936.422	90.957	.000
Grade Level	227.500	2	113.750	2.628	.073
Interaction	12.554	2	6.277	.145	.865
Residual	27871.008	644	43.278		
Total	31924.086	649	49.190		

Table 4.29a: Information on the sample of teachers by sex and grade level.

Grade level	Male	Female	Total
Primary	334 (44%)	420(56%)	754(58%)
Intermediate	76(49%)	80(51%)	156(12%)
Secondary	190(49%)	200(51%)	390(30%)
Total	600(46%)	700(54%)	1300

that teacher's competency in educational measurement at the primary, the intermediate, or the secondary level does not depend on whether that teacher is male or female (see also Table 4.29a). Consequently, this hypothesis is not rejected.

Multiple Regression Analysis

The multiple regression analysis (MR) that follows was conducted to see which variables best predict Bahrain teachers' competency in educational measurement. Table 4.30 displays the results of a stepwise multiple regression analysis for the teachers. In this regression analysis, the total score on the TEST was regressed on 10 variables; namely, number of years in teaching, measurement coursework completed, statistics coursework completed, gender, percentage of professional time devoted to classroom testing, perceived level of competency in educational measurement, perceived level of competency in test construction, major subject of teaching/ supervision, preparing a test plan, and conducting item analysis. The results revealed that there are five variables that appear to be the best predictors of teachers' competency in educational measurement. The five variables are gender, preparing a test plan, statistics coursework completed, percentage of professional time devoted to classroom testing, and number of years in teaching. These variables together account for 20

percent of the total variance in the TEST scores. It is interesting to discover that "measurement coursework completed" variable was not found among the variables that best predict the competency in educational measurement!

Table 4.30: Multiple Regression Analysis for the TEST scores with measurement preparation, and background variables for the teachers.

CORRELATIONS

STEP	TEST (score)	1	2	3	4	5	B	ß	R^2 (in step)	R^2 (change)
1	.35						5.02	.35	.12	.12*
2	-.27	-.11					-3.95	-.23	.18	.05*
3	.11	-.01	-.02				.66	.11	.19	.01*
4	.15	.17	.01	.02			.88	.10	.20	.01*
5	-.23	-.39	.17	-.06	.03		-.08	-.08	.20	.01*

Teachers: 1 = Gender, 2 = Test Plan, 3 = Statistics coursework, 4 = Professional time devoted to classroom testing, 5 = Number of years in teaching.

Summary

This chapter included a review of the study's findings, questionnaire results, TEST results, Pearson correlation results, ANOVA results, t-test results, and multiple regression (MR) analysis results. The questionnaire results reported in this chapter summarized the measurement preparation and classroom testing activities. The TEST was reliable for the total sample and for each group. Information from the TEST results indicate that the subjects in each group independently have moderate competency in

educational measurement, except for the teachers who scored below average on the TEST. By and large, the subjects appear to be most deficient in the areas of "types of tests" and "standard error of measurement". In general, correlational, and ANOVA results revealed slight, but significant, relationships between competency in measurement and several classroom testing practices variables, measurement preparation variables, and some background variables of the subjects. Multiple regression analysis results showed that the statistics coursework variable was among the variables that best predicted competency in educational measurement, whereas the measurement coursework variable was not among those variables that best predicted competency in educational measurement. Table 4.31 below shows a summary of the major findings in the study.

Table 4.31: Summary of the major findings in the study.

HO #	Stat. Test	X(S)	Group(s)	Sig.	Non-Sig.	ω^2
3	1-ANOVA	1	All	x		.18
4	1-ANOVA	2	T	x		.07
			CS	x		.24
			PT	x		.09
5	1-ANOVA	3	T	x		.04
			CS	x		.16
			PT		x	
6	1-ANOVA	4	T	x		.06
			CS		x	
			PT		x	
7	1-ANOVA	5	T	x		.01
			CS		x	
			PT		x	
8	2-ANOVA	6	All	x		
9	2-ANOVA	7	All		x	
10	t-test	8	T	x		
11	1-ANOVA	9	T		x	
12	2-ANOVA	10	T		x	

Note: HO= Null Hypothesis.; X(S)= Independent variable(s).
Independent variables: 1= Type of respondent, 2=Measurement coursework, 3=Statistics coursework, 4=Test plan, 5=Item analysis, 6=Gender by type of respondent, 7=Age by type of respondent, 8=Gender, 9=Grade level, 10=Gender by grade level.
Sig=Significant, Non-Sig=Non-significant.

T= Teachers, CS= Curriculum Specialists, PT= Prospective Teachers

CHAPTER V

SUMMARY, DISCUSSION, CONCLUSIONS, AND SUGGESTIONS

Summary of Purposes and Procedures

The purpose of this study was twofold: first, to assess the competency of Bahrain teachers, curriculum specialists, and prospective teachers in the principles of educational measurement, and second, to highlight the strengths and weaknesses among the three groups in specific areas in educational measurement, such as item analysis, score interpretation, reliability, and validity.

Two research instruments were used to accomplish the purposes of the study, a measurement test and a questionnaire. The measurement test consisted of 65 multiple-choice items. Fifty items of the measurement test were taken from Boothroyd (1990), with slight modifications made to some items to make their context familiar to Bahrain teachers, curriculum specialists and prospective teachers. The remaining 15 items of the measurement test were based on measurement criteria addressed in measurement textbooks such as Mehrens and Lehmann (1991) and Wiersma and Jurs (1990).

The questionnaire consisted of 23 items, some of which were based on prior research findings (e.g., Boothroyd, 1990; Newman, 1981; & Mayo, 1967) that showed that these variables are valid and related in one way or

another to competency in educational measurement. The questionnaire included items on classroom testing practices, measurement and statistics preparation, practical needs in educational measurement, and background variables.

The two research instruments were administered to 1300 in-service teachers, 146 curriculum specialists, and 221 prospective teachers. The in-service teachers were randomly sampled by strata (primary, intermediate, and secondary grade levels). The 146 curriculum specialists and 221 prospective teachers are the populations of these two groups. The response rate was 82 percent for the teachers, 67 percent for the curriculum specialists, and 52 percent for the prospective teachers.

Discussion and Conclusions

By and large, the teachers, curriculum specialists, and prospective teachers had taken a reasonable amount of measurement and statistics coursework. In general, fewer statistics courses were completed compared to measurement coursework. Most of the prospective teachers (80 percent) reported that they completed one college course in measurement, but only 35.6 percent of the teachers, and less than one third of the curriculum specialists (26.5 percent), said they took one college course in educational measurement. The curriculum specialists had more multiple measurement coursework than the other groups, with 36.7 percent of them reporting they had more than one college course in educational measurement. Regarding the statistics coursework completed, the picture is somewhat different. Less than one- half (43.5 percent) of the prospective teachers, 34.7 percent of the curriculum specialists, and 21.7 percent of the teachers reported that they

had one course in statistics. Furthermore, 28.6 percent of the curriculum specialists, 18.3 percent of the prospective teachers, and 5.1 percent of the teachers indicated that they had more than one course in statistics. Moreover, in-service training in both measurement and statistics was sparse, especially for statistics. Those who rated themselves as having "very good" or "good" competency in educational measurement were between 21.7 to 42.2 percent across the three groups . In addition, 21.5- 46.1 percent of the respondents across the three groups indicated that they were "very good" or "good" in test construction. Whereas 24.1 percent of the teachers and 24.5 percent of the curriculum specialists reported that they are "excellent" in test construction, less than 5 percent of the prospective teachers felt this confident. Except for the prospective teachers (for the statistics variable only), those with measurement and statistics backgrounds scored significantly higher on the TEST than those with no measurement or statistics background at all, though the proportion of variance in the TEST scores accounted for by measurement and statistics variables was small or negligible (see Tables 4.21a, 4.21b, 4.22a and 4.22b).

Unfortunately, the amount of measurement and statistics coursework completed have not translated into high performance levels for the three groups on the TEST. The curriculum specialists and the prospective teachers, in general, were more competent in measurement than the teachers. The teachers had a mean score of 26.72 (41 percent), the curriculum specialists had a mean score of 37.49 (58 percent), and the prospective teachers had a mean score of 34.07 percent (52 percent). The scores on the TEST ranged from 9 to 48 for the teachers, from 19 to 50 for the curriculum specialists, and from 18 to 49 for the prospective teachers. The teachers had the lowest average score among the three groups (see

Tables 4.20a and 4.20b). The teachers performed better in the areas of "grading and marking", and "objectives" than they did in all the other areas that comprised the TEST. The teachers were weak in many areas such as "item analysis", "standard error of measurement" and "score interpretation". The curriculum specialists performed better in the areas of "objectives", "grading and marking", and "validity" than they did in some other areas such as "test planning", "types of items", "item writing", "test construction", "item analysis", and "score interpretation." They were weak in the areas of "standard error of measurement", and "types of tests." The curriculum specialists outperformed the teachers and prospective teachers in all areas of the TEST, except for the areas of "types of items" and "reliability" (see Table 4.13). Concerning the prospective teachers, they scored better in the areas of "objectives", "types of items", "item writing", "grading and marking", "correlation", "reliability", and "validity" than they did in some other areas such as "test construction", "item analysis", "score interpretation", and "standard error of measurement."

Generally speaking, even though most of the respondents, especially the teachers and prospective teachers, reported that they conduct item analysis on their self-constructed classroom tests, this did not translate into high performance in the area of "item analysis" (see Table 4.13).

The three groups' weakness in some areas, such as "item analysis", may be reflected in their stated needs for practical training in educational measurement. The results indicated that the teachers may need more practical training in item analysis, test planning, and reliability. The curriculum specialists and prospective teachers may need more practical training in item analysis and score interpretation.

The female teachers scored significantly higher on the TEST than their male counterparts which indicates that female teachers are more competent in educational measurement than male teachers. Except for the teachers, no significant correlations were found between the TEST scores and "classroom testing" variables, and "measurement preparation" variables. Only for the teachers was there a significant correlation between the TEST scores and such variables as the purposes of using classroom tests, developing a test plan, conducting item analysis, and self-assessment of the general competency in educational measurement. Even so, the proportion of variance in the TEST scores accounted for by each of these variables was small or negligible.

It was found, for the teachers only, that those who stated that they conducted item analysis and prepared a test plan scored significantly higher on the TEST than those who did not conduct item analysis or prepare a test plan. There was no significant difference among teachers in measurement competency from one grade level to another.

The findings of this study revealed that there was an interaction between the type of respondent and respondent's gender with respect to measurement competency. The age of respondent, however, has no bearing on measurement competency from one group to another. This applies to all groups. For the teachers only, the results showed that teacher's gender has no bearing on measurement competency in the primary, intermediate, and secondary grade levels. The results of a stepwise multiple regression analysis revealed that there were five variables (gender, preparing a test plan, statistics coursework completed, amount of professional time devoted to classroom testing, and number of years in teaching) that best predict

teachers' competency in educational measurement. Interestingly, the amount of measurement coursework taken was not among them!

In this study, the curriculum specialists had more familiarity with educational measurement terms and principles than the teachers and prospective teachers. Although the teachers with measurement training scored significantly higher on the TEST than those teachers with no measurement training at all, the entire group of teachers demonstrated below average performance on the TEST.

It can be argued that in-service training in educational measurement for the teachers and curriculum specialists is minimal compared to the measurement coursework completed in a college of education.

The teachers, curriculum specialists, and prospective teachers tended to give themselves high ratings concerning self-assessed competency in educational measurement, test planning, and item analysis. These ratings, however, did not match their actual performance on the TEST (see Table 4.13).

Based on the findings of this study, the following conclusions can be drawn:

1. The teachers in this study were the least competent group in educational measurement . They need more measurement training in general, and specific training in the areas where they were weak.

2. The curriculum specialists showed that they were more competent in educational measurement than the teachers and prospective teachers, and that they indeed had a good grasp of some measurement terms in some areas (see Table 4.13).

3. The self-rating on some variables such as preparing a test plan, conducting item analysis, and self-assessment of the general competency in educational measurement, especially for the teachers and prospective teachers, did not translate into good performance on the TEST.

4. Female teachers were more competent in educational measurement than male teachers.

5. The competency in educational measurement, however, did not depend on the grade level.

6. Competency in educational measurement among the three groups depended on whether the respondent in that group is male or female; but the age of respondent had no bearing on the competency of that respondent from one group to another.

7. The "amount of measurement coursework completed" was related to measurement competency. This variable, however, was not found among those variables that best predicted competency in educational measurement.

RECOMMENDATION

The Bahrain teachers, curriculum specialists, and the prospective teachers need more training in educational measurement in general, and in the areas where they manifested around or below average performance in particular. A measurement training "liaison officer" for each public school district may assist with measurement training in schools, and may facilitate coordination and direct contact between the school(s) and the measurement specialists in the Bahrain Ministry of Education. The amount and type of

measurement training to be provided to the prospective teachers should be based on further research.

SUGGESTIONS FOR FURTHER RESEARCH

Reviewing the previously conducted research studies on teachers' competency in educational measurement, starting with Mayo's (1967) study and ending with the most recent studies in the field (e.g., Plake, Impara, & Fager, 1992), reveals some shortcomings in how this issue was approached. First, many studies used almost the same type of instruments to assess competency of teachers in educational measurement. Second, some of the questionnaires used in some research studies to assess the perceived competency of teachers and other educators in educational measurement did not appear to be revealing the actual level of measurement competency of these groups, especially when the performance on the measurement competency test was not comparable with the perceived measurement competency. Third, those studies inadvertently ignored other factors that may have some impact on, for instance, teachers' competency in educational measurement. These factors include how educational measurement was taught to teachers, the way they learned it either in a college of education or through in-service training, how their learning in educational measurement was evaluated, whether they practiced in their schools what they learned in the colleges of education, and whether the colleges of education evaluated the impact their measurement courses may have had on the teachers in schools.

Based on the findings of this study, and on the shortcomings in this study and in the research studies conducted earlier on teachers' and other

educators' competency in educational measurement, the following suggestions warrant:

1. A combination of more than one research instrument or more than one research approach in studying the competency of teachers and other groups in educational measurement may enable researchers to detect some factors that may affect competency in educational measurement. For example, one might use, in addition to the "regular" measurement competency test, a measurement test that asks for direct performance on some measurement topics that are very related to classroom assessment, such as item analysis, objectives, and test construction.

2. Some research methods such as observations and structured interviews with professors of educational measurement and prospective teachers may provide valuable data regarding competency in educational measurement. For example, professors of educational measurement may be observed teaching the entire measurement course. Structured interviews with professors of educational measurement are essential in order to explore how these professors view students' learning in educational measurement, the major strengths and weaknesses in understanding and applying measurement principles, the major factors these professors found as impediments to students' practical application of some measurement principles, and what factors among these still persist. Furthermore, interviewing a representative sample of students who took a measurement course(s) may shed light on some factors that impede improving measurement competency. These factors may include how students were introduced to educational measurement courses, what measurement principles they applied, what topics they found to be of direct link to their

daily classroom assessment, and what made them eliminate application of some measurement principles in their classrooms (e.g., item analysis).

3. A carefully designed continuing evaluation of the quality of measurement courses offered in teacher-training programs and through in-service training may be needed.

REFLECTIONS

The following are some reflections on some findings in the study and their cultural/ social implications.

1. Although there were significant gender differences in measurement competency, there were no significant grade level differences. The implication here is that there is a cultural/ social assumption/ belief that female teachers are more committed to their learning and professional development, and more organized than their male counterparts. Finding non-significant differences among the teachers at different grade levels seems contradictory with the cultural/ social assumption/ belief that those teachers who teach at the higher grade levels (e.g., secondary) are more knowledgeable, in general, than those teachers who teach at the lower grade levels (e.g., primary). There are two implications here. First, it is not necessarily true that the teachers at the higher grade levels are more knowledgeable in their subject matter than those teachers at the lower grade levels. Second, teachers may differ in their knowledge of the subject matter because they are teaching at different grade levels (e.g., a primary teacher does not need as broad an understanding in mathematics, for example, as does a secondary teacher) rather than because of the amount of measurement coursework taken.

2. When one looks at the distribution of performance on the various content areas of the TEST (Table 4.13) it is readily evident that if the last four areas of the TEST (correlation, reliability, standard error, and validity) were dropped, the TEST sores of the teachers would increase and be comparable to those of the prospective teachers. This might be attributable to the fact that because these four areas are more technical, they are less familiar to the practicing teachers than to the "newer" prospective teachers. In addition, the difference might be due to better training today in these areas. Furthermore, since knowledge of statistics appeared to be a good predictor of measurement competency, this might account for the difference.

3. The difference in the self-ratings between the teachers and prospective teachers (see Tables 4.9 and 4.10) can be looked at from different angles. The large sample size of the teachers may have a bearing on that difference. Second, the teachers may be more honest in their ratings than the prospective teachers, because the teachers feel more responsible than the prospective teachers and they take the issue more seriously. Third, the prosepctive teachers, however, may have thought that their knowledge in measurement is more recent, and hence that they are competent. Finally, teachers' experience with test construction and classroom tests may have and influence on their self-ratings when compared with the prospective teachers who almost lack that experience.

APPENDIX A

TEST OF EDUCATIONAL MEASUREMENT

DIRECTIONS

1.This test consists of 65 questions.

2.In the space provided to the left of each item, please print the letter of the option which you believe BEST answers the question.

3.Please print only ONE letter for each question.

4.You have 3 HOURS to complete the test (you may take extra time if you need).

5.Please answer all questions.

****

___1. TABLES OF SPECIFICATIONS are important because they help insure the tests'
a.content validity.
b.criterion-related validity.
c.internal consistency.
d.test-retest reliability.

___2.Essay items are more appropriate than multiple-choice items for assessing
a.application of knowledge.
b.factual knowledge.
c.recognition of ideas.
d.organization of ideas.

___3.A DISADVANTAGE of multiple choice tests as compared to essay tests is that they are
a.generally less valid.
b.scored more subjectively.
c.more time consuming to construct.
d.appropriate only for narrow content domains.

___4.A test plan (Table of Specifications) is LEAST useful in
a.judging the relative emphasis to be given to different content areas.
b.judging the relative emphasis to be given to different objectives.
c.deciding the type of items to use on the test.
d.relating the item content to performance (behavioral) objectives.

___5.The type of evidence with which a teacher should be most concerned when constructing classroom tests is
a.construct validity.
b.content validity.
c.split-half reliability.
d.test-retest reliability.

___6."chance" success in matching items can best be reduced by
a.using statements that require fine discriminations.
b.not allowing responses to be used more than once.
c.lengthening the statements and the responses.
d.including more responses than statements.

___7.Which one of the following is NOT considered a sound principle in constructing true-false questions?
a.Avoiding the use of negative statements as much as possible.
b.Confining each question to a single idea.
c.Using more true than false statements.
d.Refraining from using long and complex statements.

___8.Instructional objectives are MOST useful when they
a.are equally appropriate for all students in the class.
b.are specified in terms of the knowledge they involve.
c.relate skills and knowledge to performance (behavioral) objectives.
d.can be used as the basis for developing objective tests.

___9.For a test to be valid it must also be
a.practical.
b.reliable.
c.important.
d.objective.

___10.A criterion-referenced interpretation is MORE appropriate than a norm-referenced interpretation for decisions concerning
a.classroom strengths and weaknesses.
b.determination of mastery.
c.placement in a gifted program.
d.prediction of student performance.

___11.Which one of the following item types is BEST suited to measure a student's ability to apply information?
a.Matching items.
b.Multiple-choice items.
c.True-false items.
d.Both b and c.

___12."Manama is the most important city in Bahrain" is a poor true-false item because it is
a.ambiguous.
b.too easy.
c.too factual.
d.trivial.

___13.Which one of the following item types is LEAST objective?
a.Matching items.
b.Completion items.
c.Multiple-choice items.
d.True-false items.

___14.Since no test is perfectly reliable, a student's score should be regarded as
a.having little meaning unless it is very high or very low.
b.having little validity in estimating the student's achievement.
c.indicating a point in a range of scores in which the student's "true score" probably falls.
d.indicating only that the student has more or less ability than the average student.

___15.In the scoring of an historical essay test, all of the following are generally considered desirable EXCEPT
a.preparing a scoring key and standards in advance.
b.deducing points for grammatical errors and poor handwriting.
c.removing the students' names from the test.
d.scoring one question on each test prior to going on to the next.

___16.The practice of allowing students a choice in the questions to be answered on an essay exam is generally considered to be
a.undesirable, because students waste too much time deciding which questions to answer.
b.undesirable, because it reduces the comparability of the test from student to student.
c.desirable, because it gives each student a fairer time.
d.desirable, because it permits a wider sampling of topics covered.

___17.Arranging test items by item type tends to
a.make test taking easier for the student.
b.reduce the standard error of measurement.
c.increase item validity.
d.increase the test's reliability.

___18.A norm-referenced interpretation compares each examinee's score to
a.a performance standard.
b.other examinees' performance.
c.the true score of the individual.
d.the mean test score.

___19.Layla attained a score of 80 on her criterion-referenced geography test. Which of the following additional piece of information is MOST useful for interpreting her performance?
a.The minimum passing score was 75.
b.The class average score was 65.
c.The test had 90 questions.
d.Two thirds of the class failed the test.

___20.If an item is correctly answered by 38% of the students in the top third of the class, and 66% of the students in the bottom third, the item's discriminating power is
a.negative.
b.unstable.
c.valid.
d.almost perfect.

___21.The correlation coefficient between two sets of scores from the same test given to the same class is .95 . This indicates that the test is highly
a.discriminating.
b.valid.
c.reliable.
d.unstable.

___22.Which correlation coefficient indicates the highest predictability of one variable from the other?
a. -.91
b. -.35
c. +.54
d. +.76

___23.The standard error of measurement is a useful index of the
a.amount of variability in the score distribution of all persons tested.
b.proportion of persons answering the item incorrectly.
c.degree of variability of a single person's observed score on multiple testings.
d.test's validity.

___24.Rashid knows absolutely nothing about the material in a 40 item, four-option, multiple-choice test. By random guessing he should answer how many items correctly?
a.0
b.4
c.10
d.20

___25.For which of the following learning outcomes would a multiple-choice test be LEAST relevant?
a.Defines a basic concept.
b.Identifies the reason for an action.
c.Relates an example to a stated principle.
d.Selects the best method to use.

___26.A sixth grade class was given two different forms of the same science test. The test scores from the two forms placed students in nearly the same order. This would be evidence concerning the test's
a.internal consistency reliability.
b.equivalence reliability.
c.scorer reliability.
d.stability reliability.

___27.A test is reliable when it
a.can be scored quite easily.
b.gives a consistent estimate of whatever it measures.
c.measures what it is designed to measure.
d.provides a basis for diagnosing pupil's weakness.

___28.Which one of the following is a characteristic of a highly reliable test?
a.Good students score higher than poor students.
b.The items in the test vary widely in difficulty.
c.Two scores for the same examinee agree with each other.
d.There is a uniform distribution of scores on the test.

___29.If the scores on a test are 5, 2, 3, 27, and 8, the test's mean score is
a.5
b.7
c.8
d.9

___30.An item with a difficulty level of 85% indicates that
a.the item has a good discrimination.
b.the item is easy.
c.some students answered the item correctly.
d.a large proportion of students who did poorly on the test got the item wrong.

___31.An item's discrimination will likely be maximized when its difficulty index approaches
a. 0%
b. 25%
c. 50%
d. 100%

___32. Which one of the following is the MOST serious problem with the true-false item: "Manama is the capital city of Bahrain, a major banking center, and an important port".
a. Tests multiple concepts.
b. Too many specific determinants.
c. Too easy.
d. Contains grammatical errors.

___33. In terms of Bloom's taxonomy, application reflects a higher cognitive function than does
a. comprehension.
b. analysis.
c. evaluation.
d. synthesis.

___34. "The student will demonstrate comprehension of the short story form" is a poor instructional objective because it
a. is too unrealistic.
b. describes the instructional process.
c. does not specify the conditions under which the performance is expected.
d. does not specify an observable performance.

___35. Which one of the following item types would be MOST appropriate for determining students' ability to compose a short story?
a. Essay items.
b. Completion items.
c. Multiple-choice items.
d. Short-answer items.

___36. If Ahmad had a test score of 78 on a test with an estimated standard error of 6, we would be 95% confident that his "true score" would fall between
a. 66 and 90.
b. 69 and 87.
c. 72 and 84.
d. 75 and 81.

___37.A percentile rank refers to the percentage of
a.items correctly answered.
b.items needing to be correctly answered to obtain a passing score.
c.students scoring above the mean on the test.
d.students scoring below a given raw score.

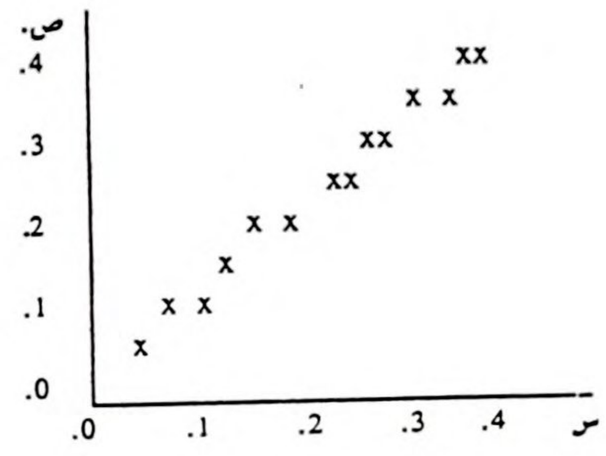
___38.An advantage short-answer items have over multiple-choice items is that **they**
a.are more reliable.
b.can be used on either norm-referenced or criterion-referenced tests.
c.have more desirable discrimination levels.
d.minimize random guessing.

___39.Which one of the following statements MOST suggests a norm-referenced interpretation?
a.Noora scored higher than the class average score on the science test.
b.Mariam correctly answered 80% of the items on the mathematics test.
c.Hassan passed the minimum competency reading test.
d.Ali exceeded the mastery level on the 5th grade social studies.

___40.A history teacher uses the following item: "List three reasons why the Gulf war started". To classify this question on Bloom's taxonomy we would need to know
a.what the students received for Gulf war lectures and readings.
b.the general intellectual level of the students in the class.
c.what the teacher keyed as acceptable answers to the questions.
d.whether the test is norm-referenced or criterion-referenced.

___41.A correlation coefficient indicates the
a.extent of variation within a set of test scores.
b.extent to which two variables measure the same thing.
c.degree to which one variable causes the other.
d.degree of relationship between two variables.

___42.



The relationship indicated by the above scatter plot is
a.highly negative.
b.moderately negative.
c.moderately positive.
d.highly positive.

___43.Which one of the following is appropriate for assessing a test's internal consistency?
a.Equivalent forms.
b.Kuder-Richardson
c.Coefficient of stability.
d.Test-retest.

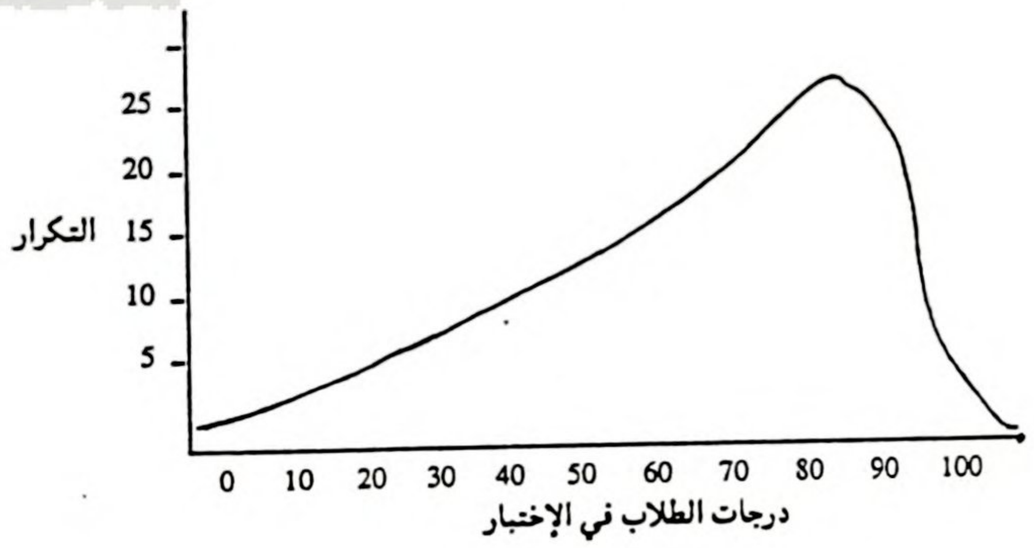
___44.A <u>DISADVANTAGE</u> of a norm-referenced grading scheme is that
a.percentage grades cannot be calculated.
b.an individual's class ranking cannot be determined.
c.grading is presented in relative terms.
d.performance is assessed in terms of a fixed criterion.

___45. Which one of the following statements BEST illustrates a criterion-referenced interpretation?
a. Nada earned the highest score in the class.
b. Amal set up her experiment in five minutes.
c. Khalid's test score was average for his class.
d. Ebrahim solved the fewest arithmetic problems.

___46. An important consideration when constructing classroom tests is to
a. construct items which require students to make fine distinctions.
b. develop two parallel forms of the test.
c. match the items to instructional objectives.
d. test materials not taught in class so that students have an opportunity to demonstrate their out-of-class learning.

___47. Which one of the following item statistics would be of MOST serious concern to the teacher?
a. Difficulty index= .80
b. Difficulty index= .05
c. Discrimination index= .00
d. Discrimination index= -.05

___48.



The graph above indicates that the test is
a.too easy.
b.too difficult.
c.moderately difficult.
d.moderately easy.

___49.Objective scoring of selected-response items enhances the test's
a.reliability.
b.validity.
c.validity but not stability.
d.reliability and validity.

___50.If the reliability of a test is 1.00, the standard error of measurement is
a. .00
b. .10
c. .50
d. 1.00

___51.The point below which 50% of the scores lie is known as the
a.median.
b.mean.
c.mode.
d.standard deviation.

___52.Which one of the following is a measure of variability of scores?
a.Mean.
b.Median.
c.Mode.
d.Standard deviation.

___53.According to Bloom's taxonomy, at what level can the following statement be classified. "The teacher said to one of his students: Your composition passage is of poor quality"?
a.Application.
b.Analysis.
c.Synthesis.
d.Evaluation.

___54.From a measurement viewdpoint, using classroom tests consisting entirely of essay items is undesirable because
a.scoring requires too much time.
b.it is difficult to construct the items.
c.content sampling tends to be limited.
d.constructing model responses is too time consuming.

___55.Which one of the following is NOT generally associated with criterion-referenced testing?
a.A performance standard.
b.A well specified content domain.
c.Test items selected according to difficulty level.
d.The use of test items which closely reflects the instructional objectives.

___56.A poorly defined domain results in items that are
a.too difficult.
b.dissimilar.
c.ambiguous to the student.
d.unreliable.

___57.The term objective , when used to label educational tests, MOST appropriately describes
a.a characteristic of the scoring procedure.
b.the degree of the standardization of the test.
c.the content limitation of the question.
d.whether the test interpretation is norm-referenced or criterion-referenced.

___58. Which one of the following would indicate the highest level of learning for a student?
The student
a. applies a principle.
b. gives a precise definition of the principle.
c. identifies an example using the principle.
d. rephrases a principle in his/her own words.

___59. Which one of the following educational goals is NOT stated in performance (behavioral) terms?
a. Design.
b. Prepare.
c. Understand.
d. Write.

___60. A major strength of criterion-referenced testing is that one can
a. generalize the results to a larger set of items.
b. compare students in terms of relative standing.
c. establish objective performance criteria.
d. measure a specific trait.

___61. Problems arise when attempting to develop measures of ultimate goals mainly because
a. group norms with which to compare the results are generally not available.
b. measurement methods have not given proper weight to all goals.
c. such goals concern behaviors usually not observable under classroom conditions.
d. teachers are reluctant to depart from traditional testing methods.

___62. Which one of the following educational outcomes is classified in the affective domain?
a. Developing handwriting skills.
b. Fostering an interest in music.
c. Improving reading comprehension.
d. Understanding the classification of interests.

___63. What is the major problem with the following completion item in a seventh grade history test: "The Portuguese invaded Bahrain in_____"?
a. Too easy.
b. Unclear.
c. Has a specific determiner.
d. measures factual knowledge

___64. When constructing multiple-choice tests, it is best to
a. make all options the same length.
b. put the main idea of the item in the options.
c. use options such as a and b , but not c.
d. repeat key words from the stem in the options.

___65. Completion and short answer items should have the blank(s) placed
a. at or near the beginning of the item.
b. between the beginning and the middle of the item
c. as close to the middle of the item as possible.
d. at or near the end of the item.

****

THANK YOU VERY MUCH FOR YOUR COOPERATION

THE QUESTIONNAIRE

DIRECTIONS.

1.This questionnaire consists of 23 questions.
2.Please use the space(s) provided to print your response(s).
3.Please note that some questions require that you select only ONE number from among the numbers that correspond to the options provided, and print that number in the space provided.
4.Please note that in some questions you can select more than one option.
5.Please answer all questions.

****

1.___ Age?

2.___ Gender?
[1] Male
[2] Female

3.___ Education level (the highest degree you hold). Year obtained _____.
[1] Ph.D.
[2] Masters
[3] High Diploma (beyond bachelors degree)
[4] Bachelors
[5] Associate Degree (Diploma) (two years beyond high school)
[6] High School Certificate
[7] Less than High School Certificate

4. Is the highest degree you indicated in Question 3 above obtained from a college of education? ___ Yes. ___ No.
If your answer is NO, please specify the highest degree you obtained from a college of education. _____.

5. Was EDUCATION your MAJOR when you were in your senior year in the undergraduate program? ___ Yes. No. ___.

6. _____ Major subject of specialization?

7. _____ Subject you teach/ supervise.

8. ___ Stage at which you teach/ supervise?
   [1] Primary
   [2] Intermediate
   [3] Secondary

9. ___ Number of years of experience in teaching? (for in-service teachers and curriculum specialists only).

10. ___ Number of years of experience in curriculum supervision? (for curriculum specialists only).

11. How much coursework in <u>educational measurement</u> have you completed? (check as many as apply).
   __ None
   __ Part of another course.
   __ One college course.
   __ More than one college course.
   __ As in-service course/ training (as offered by the Ministry of Education).

12. How much coursework in <u>statistics</u> have you completed? (check as many as apply).
   __ None.
   __ Part of another course.
   __ One college course.
   __ More than one college course.
   __ as in-service course/ training (as offered by the Ministry of Education).

13. ___ In what year did you complete your <u>most recent</u> coursework in educational measurement?

14. ___ In what year did you complete your <u>most recent</u> coursework in statistics?

15. Approximately what percentage of your professional time (both in and out of school) do you devote to constructing, administering, scoring, grading, interpreting, and reviewing your self-constructed classroom tests? (check <u>one</u> only).
   __ Less than 5%.
   __ 5% to 10%.
   __ 11% to 20%.
   __ More than 20%.

16. For what purposes do you use your self-constructed classroom tests? (check as many as apply).

__ Diagnosing strengths and weaknesses.
__ Assessing academic achievement.
__ Assessing mastery of instructional units.
__ Assessing academic progress.
__ Motivating learning.
__ Assigning grades.
__ Planning instruction.
__ Other.

17. What type of items do you write for your self-constructed classroom tests? (check as many as apply).

__ True-False.
__ Matching.
__ Multiple-Choice.
__ Completion (fill-in-the-blanks).
__ Short Answer (one or two paragraphs).
__ Essay (one or two pages).
__ Other (specify). _____.

18. Rank the following item types in terms of their preference to you (1=Most preferred, 6=Least preferred).

__ True-False.
__ Matching.
__ Multiple-Choice.
__ Completion (fill-in-the-blanks).
__ Short Answer (one or two paragraphs).
__ Essay (one or two pages).

19. ___ Do you generally develop a test plan (e.g., table of specifications) prior to writing test items?

[1] Yes.
[2] No.

20. ___ Do you generally conduct an item analysis on your self-constructed classroom tests?

[1] Yes.
[2] No.

21. ___ How would you assess your competency in educational measurement in general?
[1] Excellent.
[2] Very good.
[3] Good.
[4] Adequate.
[5] Poor.

22. ___ How would you assess your competency in test construction principles?
[1] Excellent.
[2] Very good.
[3] Good.
[4] Adequate.
[5] Poor.

23. In which of the following topics in educational measurement you feel that you need practical training ? (choose as many as apply).
__ Planning classroom tests (e.g., table of specifications).
__ Constructing different test item types (e.g., true-false).
__ Knowing what test item type to choose and why for constructing your classroom tests.
__ Conducting item analysis on your classroom tests (e.g., item difficulty, item discrimination).
__ Test validity (e.g., content validity).
__ Test reliability (e.g., internal consistency).
__ Interpreting classroom test scores.
__ Observational methods.
__ Other (please specify). _____.

****

THANK YOU VERY MUCH FOR YOUR COOPERATION

APPENDIX B

APPENDIX B

THE TABLE OF SPECIFICATIONS OF THE TEST

Content Area	Item Level				Item Total
	Knowledge of Terms	Knowledge of facts	Explanation & Interpretation	Application of Numerical Problems	
Test Planning		1	4		2
Objectives		8, 33, 53, 59	34,40,58,61, 62		9
Types of Tests	57	18, 56, 60, 55	10		6
Types of Items		2, 11, 13, 25, 38	3, 35, 54		8
Item Writing		6, 7, 64, 65	12, 32, 63		7
Test Construction		16, 17, 46	47		4
Item Analysis	51, 52	31	30	20	5
Score Interpret-ation		37, 44, 49	14, 19, 39, 45, 48		8
Grading & Marking		15		24, 29	3
Correlation	41		22, 42		3
Reliability	26, 27, 43	28	21		5
Standard Error	23			36, 50	3
Validity		5, 9			2
Item Total	**8**	**29**	**23**	**5**	**65**

APPENDIX C

THE TEST ANSWER KEY

1. A	22. A	43. B	64. A
2. D	23. C	44. C	65. D
3. C	24. C	45. B	
4. C	25. A	46. C	
5. B	26. B	47. D	
6. D	27. B	48. A	
7. C	28. C	49. A	
8. C	29. D	50. A	
9. B	30. B	51. A	
10. B	31. C	52. D	
11. D	32. A	53. D	
12. A	33. A	54. C	
13. B	34. D	55. C	
14. C	35. A	56. B	
15. B	36. A	57. A	
16. B	37. D	58. A	
17. A	38. D	59. C	
18. B	39. A	60. A	
19. A	40. A	61. C	
20. A	41. D	62. B	
21. C	42. D	63. B	

APPENDIX D

بسم الله الرحمن الرحيم

إختبار في القياس التربوي

إرشـــــــادات

١ - يتكون هذا الإختبار من ٦٥ سؤالاً.
٢ - في الفراغ المخصص عن يمين كل سؤال أكتب من فضلك حرف الخيار الذي تعتقد أنّه يجيب عن السؤال بالشكل الأفضل.
٣ - إختر إجابه واحده فقط.
٤ - أجب من فضلك عن جميع الأسئله.
٥ - لديك ثلاث ساعات للإجابه (يمكنك اخذ وقت إضافي).

.....

ـــ ١ . جداول المواصفات للإختبارات مهمّه لأنها تساعد على ضمان
أ . صدق المحتوى.
ب . الصدق المرتبط بمعيار (بمحك)
جـ . الإتساق (التوافق) الداخلى.
د. ثبات الإختبار وإعادة الإختبار.

ـــ ٢ . الأسئلة المقاليّه أكثر مناسبة من اسئلة الإختيار من متعدد لقياس
أ . تطبيق المعرفة.
ب . معرفة الحقائق.
جـ . التعرفّ على الأفكار.
د . تنظيم الأفكار.

ـــ ٣ . أحد عيوب إختبارات الإختيار من متعدد مقارنة بالإختبارات المقاليّه هو انّها
أ . أقل صدقاً بشكل عام.
ب . يتم تصحيحها بشكل أكثر ذاتيّه.
جـ . أكثر استهلاكاً للوقت عند بنائها (إعدادها).
د. مناسبه للموضوعات (للمواد) ذات المحتوى الضيّق.

ـــ ٤ . خطّة الإختبار (جدول المواصفات) أقلّ فائدة في
أ . الحكم علي الإهتمام النسبي الّذي يُعطى لأقسام المحتوي المختلفه.
ب . الحكم على الإهتمام النسبي الّذي يُعطى لأهداف مختلفه.
جـ . تحديد نوع الأسلئه التي تُستخدم في الإختبار.
د . ربط محتوى السؤال بأهداف الأداء (الأهداف السلوكيّة).

ـــ ٥ . نوع الدليل الذي يجب أن يهتم به المدرّس بالشكل الأكبر عند بناء (إعداد) أسئلة الإختبار الصفّي هو
أ . الصدق البنائي (صدق المبنى).
ب . صدق المحتوى.
جـ . ثبات التجزئه النصفيّه.
د . ثبات الإختبار وإعادة الإختبار.

ـــ ٦ . نجاح «الصّدفة» في أسئلة المطابقة (المزاوجه) يمكن تقليله بالشكل الأفضل بواسطة
أ . استخدام عبارات تتطلّب تمييزاً دقيقاً.
ب . عدم السماح باستخدام الإجابات أكثر من مرّة.
جـ . إطالة عبارات الأسئلة والإجابات.
د . جعل عدد الإجابات أكثر من عدد عبارات الأسئلة.

ـــ **٧ . أيّ واحد من الآتي لا يعُتبر مبدأ سليماً في بناء (إعداد) أسئلة الصواب والخطأ؟**
أ . تجنب استَخدام عبارات النفي بقدر الإمكان.
ب . حصر كل سؤال في فكرة واحده.
جـ . جعل عبارات الصواب أكثر من عبارات الخطأ.
د . الإمتناع عن استخدام عبارات طويلة ومعقّده.

ـــ **٨ . تكون الأهداف التعليميّة أكثر فائدة عندما**
أ . تكون مناسبة بشكل متساو لجميع الطلاب في الصف.
ب . تكون محدّدة بالنسبة للمعرفة التي تتضمّنها.
جـ . تربط المهارات والمعرفة بأهداف الأداء.
د . يمكن استخدامها كأساس لتطوير اختبارات موضوعيّة.

ـــ **٩ . لكي يكون الإختبار صادقاً يجب أيضاً أن يكون**
أ . عملياً.
ب . ثابتاً.
جـ . مُهماً.
د . موضوعياً.

ـــ **١٠ . التفسير المعياري المرجع أكثر مناسبة من التفسير الجماعي المرجع بالنسبة للقرارات المتعلّقة**
أ . بجوانب القوّة والضعف في الصف.
ب . بتحديد مستوى الإتقان.
جـ . بإلحاق الطالب في برنامج للموهوبين.
د . بالتنبؤ بإداء الطالب.

ـــ ١١. أي واحدٍ من أنواع الأسئله مما يأتي هو الأفضل ملائمة لقياس قدرة الطالب على تطبيق المعلومات؟
أ . المطابقة (المزاوجه).
ب . الإختيار من متعددٍ.
جـ . الصواب والخطأ.
د . كلّ من ب، ج.

ـــ ١٢. «المنامه هي المدينة الأهم في البحرين». هذا السؤال (من نوع صواب وخطأ) سؤال ضعيف لأنّه
أ . غامض.
ب . سهل جداً.
جـ . واقعي جداً.
د . تافه.

ـــ ١٣. أيّ واحد من أنواع الأسئله الآتيه يعتبر الأقل موضوعيّه؟
أ . المطابقه (المزاوجه).
ب . الإكمال.
جـ . الإختيار من متعدد.
د . الصواب والخطأ.

ـــ ١٤. طالما أنّه لا يوجد اختبار ثابت تماماً، فإنّ درجة الطالب يمكن اعتبارها على أنّها
أ . ذات معنى ضئيل ما لم تكن عاليه جداً أو منخفضه جداً.
ب . ذات صدق ضئيل في تقدير تحصيل الطالب.
جـ . تشير إلى نقطة في مدى الدرجات حيث يُحتمل أن تقع «الدرجة الحقيقية» للطالب.
د . تشير فقط إلى أن الطالب لديه قدرة أكثر أو أقل من الطالب المتوسط.

ـــ ١٥. عند تصحيح اختبار مقالي في التاريخ، فإنّ كلّ مما يأتي يُعتبرُ مرغوباً بشكل عام باستثناء
أ . إعداد معايير ومفتاح للتصحيح بشكل مسبق.
ب . خصم درجات بسبب الأخطاء النحويّه ورداءة الخط.
جـ . إزالة أسماء الطلاب من أوراق الإختبار.
د . تصحيح سؤال واحد بكل أوراق الإختبار قبل الإنتقال إلى تصحيح سؤال آخر.

ـــ ١٦. إنّ السماح للطلاب بالإختيار في أسئلة الإمتحان المقالي يُعتبرُ بشكل عام
أ . غير مرغوب لأن الطلاب يضيعون الكثير من الوقت في تحديد أي سؤال يجب أن يجيبوا عنه.
ب . غير مرغوب لأنّه يُقلّل من قابليّة المقارنه في الإختبار من طالب إلى آخر.
جـ . مرغوب لأنّه يمنح كل طالب فرصة أكثر عدلاً.
د . مرغوب لأنّه يسمح بتغطية عينّه أكبر من الموضوعات التي تمّت دراستها.

ـــ ١٧. ترتيب أسئلة الإختبار حسب نوع السؤال يميل إلى
أ . جعل عمليّة أخذ الإختبار أسهل بالنسبة للطالب.
ب . تقليل الخطأ المعياري للقياس.
جـ . زيادة صدق السؤال.
د . زيادة ثبات الإختبار.

ـــ ١٨. التفسير الجماعي المرجع يقارن درجة كل طالب مع
أ . مستوى (معيار) للأداء.
ب . أداء الطلاب الآخرين.
جـ . الدرجة الحقيقية للطالب.
د . متوسّط درجة الإختبار.

ـــ ١٩. حصلت ليلى على درجة ٨٠ في اختبار الجغرافيا المعياري المرجع. أيّ من المعلومات الإضافيّه الآتيه يُعتبر الأكثر فائدة في تفسير أدائها
أ . درجة الحد الأدنى للنجاح كانت ٧٥.
ب . متوسّط درجة الإختبار للصف كان ٦٥.
جـ . كان الإختبارُ مكوّناً من ٩٠ سؤالاً.
د . ثلثا الطلاّب بالصف رسبوا في الإختبار.

ـــ ٢٠. إذا أجاب ٣٨٪ من الطلاّب الّذين هم في الثلث الأعلى بالصف، وكذلك أجاب ٦٦٪ من الطلاب الّذين هم في الثلث الأسفل بالصف، عن أحد الأسئلة إجابة صحيحه؛ فإنّ القوةّ التمييزيّه للسؤال تكون
أ . سالبه.
ب . غير مستقرّة.
جـ . صادقه.
د . تامّه تقريباً.

ـــ ٢١. معامل الإرتباط بين مجموعتين من الدرجات لنفس الإختبار ولنفس الصف هو ٩٥و. ، يشير ذلك إلى انّ الإختبار عالي
أ . التمييز.
ب . الصدق.
جـ . الثبات.
د . في عدم الإستقرار.

ـــ ٢٢.أيّ من معاملات الإرتباط الآتيه يشير إلي أعلى قابليّة للتنبؤ من متغيّر إلى آخر؟
أ . -٩١و.
ب . -٣٥و.
جـ . ٥٤و.
د . ٧٦و.

ـــ ٢٣. يُعتبر الخطأ المعياري للقياس مؤشراً ذو فائده في معرفة
أ . مقدار التباين في توزيع الدرجات لجميع الأشخاص الّذين تمّ اختبارهم.
ب . نسبة الأشخاص الّذين أجابوا عن السؤال إجابة خاطئه.
جـ . مقدار التباين في درجة شخص واحد في اختبارات متعدّده.
د . صدق الإختبار.

ـــ ٢٤. راشد لا يعرف أي شيء إطلاقاً عن مضمون مادّة الإختبار في اختبار من نوع الإختيار من متعدّد ذو الأربعة خيارات، ومكوّن من ٤٠ سؤالاً. كم سؤالاً يستطيع أن يجيب عنها إجابه صحيحه عن طريق التخمين العشوائي ؟
أ . صفر
ب . ٤
جـ . ١٠
د . ٢٠

ـــ ٢٥. لأي من نواتج (نتائج) التعلّم الآتيه يكون إختبار من نوع الإختيار من متعدّد أقل ملائمة؟
أ . تعريف مفهوم أساسي.
ب . التعرّف على السبب في القيام بعمل ما.
جـ . ربط مثال بمبدأ مُحدّد.
د . إختيار الطريقة الأفضل لاستخدامها.

ـــ ٢٦. تمّ إعطاء نموذجين مختلفين من نفس اختبار العلوم للصف السادس الإبتدائي. درجات الإختبار في كلا النموذجين وضعت الطلاب في نفس الترتيب تقريباً. بالنسبة للإختبار، يُعتبر ذلك دليلاً مرتبطاً بثبات
أ . الإتساق (التوافق) الداخلي.
ب . التكافؤ.
جـ . مُقدّر الدرجات.
د . الإستقرار.

ـــ ٢٧. يكون الإختبار ثابتاً عندما
أ . يمكن تقدير درجاته بسهولة تامّة.
ب . يُعطي تقديراً مُتّسقاً (متوافقاً) لأي شيء يقيسه.
جـ . يقيس ما هو مُصَمّمٌ لقياسه.
د . يضع أساساً لتشخيص ضعف التلميذ.

ـــ ٢٨. أيّ من الآتي يُعتبرُ خاصّية لإختبار عالي الثبات؟
أ . الطلبة الجيّدون حصلوا على درجات أعلى من الطلبة الضعفاء.
ب . أسئلة الإختبار تتباين في صعوبتها بشكل كبير.
جـ . درجتان لنفس الطالب تتوافقان مع بعضهما.
د . هناك توزيع متماثل لدرجات الإختبار.

٢٩. إذا كانت الدرجات في اختبارها ما هي ٥، ٢، ٣، ٢٧، ٨ ؛ فإنّ متوسّط درجات الإختبار هو
أ . ٥
ب . ٧
جـ . ٨
د . ٩

ـــ ٣٠. السؤال الذي مستوى صعوبته ٨٥و٠، يشير إلى أنّ
أ . السؤال ذو تمييز جيّد.
ب . السؤال سهل.
جـ . بعض الطلاب أجابوا عن السؤال إجابة صحيحه.
د . نسبة كبيره من الطلاب الّذين كان أداؤهم ضعيفاً في الإختبار أجابوا عن السؤال إجابةٍ خاطئه.

ـــ ٣١. من المحتمل أن يصل مستوى تمييز السؤال إلى حدّه الأعظم عندما يقترب معامل الصعوبه من
أ . صفر٪
ب . ٢٥٪
جـ . ٥٠٪
د . ١٠٠٪

ـــ ٣٢. أي من الآتي يُعتبرُ المشكلة الكبرى في سؤال الصواب والخطأ الآتي:
«المنامه عاصمة البحرين، وهي مركز مصرفي كبير، وميناء هام»؟
أ . يختبر مفاهيم متعدّده.
ب . مُحدِّدات الإجابه كثيره جداً.
جـ . سهل جداً.
د . يحتوي على أخطاء نحويّه.

ـــ ٣٣. بناءً على تصنيف بلوم، فإنّ التطبيق يعكس وظيفة معرفيّة أعلى من تلك التي يعكسها
أ . الفهم.
ب . التحليل.
جـ . التقويم (التقييم).
د . التركيب.

ـــ ٣٤. «سوف يُظهرُ الطالبُ فهماً لصيغة القصه القصيرة». يُعتبرُ ذلك هدفاً تعليمياً ضعيفاً لأنه.
أ . غير واقعي جداً.
ب . يصف العمليّة التعليميّه.
جـ . لا يحدّد الظروف التي يُتَوقّعُ فيها الأداء.
د . لايحدّد أداءً يمكن ملاحظته (مشاهدته).

ـــ ٣٥. أي نوع من أنواع الأسئلة الآتيه يُعتبرُ الأكثر ملائمةً لتحديد قدرة الطالب على إنشاء قصه قصيره؟
أ . المقاليّه.
ب . الإكمال.
جـ . الإختيار من متعدّد.
د . ذات الإجابة القصيره.

ـــ ٣٦. إذا حصل أحمد على الدرجه ٧٨ في اختبار إنحرافه المعياري هو ٦؛ عندئذٍ نكون واثقين بنسبة ٩٥٪ بأنّ «درجته الحقيقيّه» تقع بين
أ . ٦٦ و ٩٠
ب . ٦٩ و ٨٧
جـ . ٧٥ و ٨١
د . ٧٢ و ٨٤

ـــ ٣٧. الرتبه المئينيّه تشير إلى النسبه المئويّه
أ . للأسئلة التي أجيبَ عنها بشكل صحيح.
ب . للأسئلة التي يتعيّن الإجابة عنها بشكل صحيح للحصول على درجة النجاح.
جـ . للطلاب الّذين حصلوا على درجات أعلى من متوسط درجة الإختبار.
د . للطلاب الّذين حصلوا على درجات أدنى من إحدى الدرجات الخام.

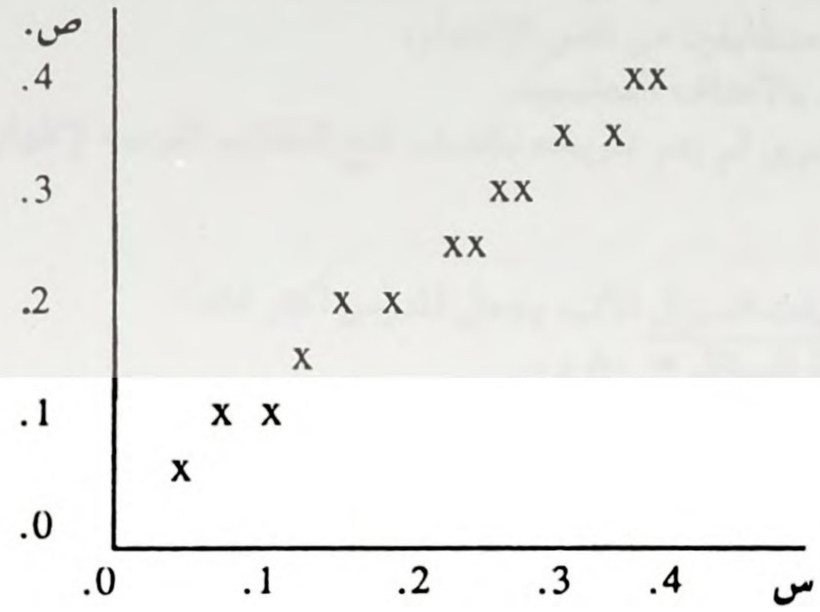
ـــ ٣٨. إحدى الميزات التي تتفوق بها أسئلة الإجابه القصيرة على أسئلة الإختيار من متعدّد هي
أ . أنّها أكثر ثباتاً
ب . أنّه يمكن استخدامها في أي من الإختبارات الجماعيّة المرجع أو المعياريّة المرجع.
جـ . أنّ بها مستويات تمييز مرغوبه بشكل أكثر.
د . أنّها تقلل من التخمين العشوائي إلى الحد الأدنى.

ـــ ٣٩. أي من العبارات الآتيه هي الأكثر إيحاءً بتفسير جماعي المرجع؟
أ . حصلت نوره على درجه أعلى من متوسط درجة الصف في اختبار العلوم.
ب . أجابت مريم إجابةً صحيحة عن ٨٠٪ من الأسئله في اختبار الرياضيات.
جـ . إجتاز حسن إختبار الحد الأدنى للكفاءه في القراءه.
د . تخطّى علي مستوى الإتقان في المواد الإجتماعيّه للصف الخامس الإبتدائي.

ـــ ٤٠. يستخدم مدرّس التاريخ السؤال الآتي: «أذكر ثلاثة أسباب أدّت إلى بدء حرب الخليج». لكي يتم تصنيف هذا السؤال ضمن تصنيف بلوم، نحتاج أن نعرف
أ . ما تلقاه الطلاب من محاضرات وقراءات عن حرب الخليج.
ب . المستوى الفكري العام للطلاب بالصف.
جـ . ما حدّده المدّرس كمفتاح للإجابة المقبوله عن الأسئله.
د . ما إذا كان الإختبار جماعي المرجع أو معياري المرجع.

ـــ ٤١. يشير معامل الإرتباط إلى
أ . مدى التباين في مجموعة درجات إختبار ما.
ب . المدى الذي يقيس فيه متغيّران نفس الشيء.
جـ . المستوى الذي يسبب فيه متغيّر حدوث متغيّر آخر.
د . درجة (مستوى) العلاقه بين متغيّرين.

— ٤٢.



العلاقة المشار إليها بواسطة الرسم البياني أعلاه
أ . سالبه بدرجة عاليه.
ب . سالبه بدرجةَ متوسطه.
جـ . موجبه بدرجةً متوسطه.
د . موجبه بدرجةٍ عاليه.

— ٤٣. أي من الأساليب الآتيه المستخدمه لحساب ثبات الإختبار هو المناسب لقياس الإتساق (التوافق) الداخلي للإختبار؟
أ . النماذج المتكافئه (من نفس الإختبار).
ب . كودر ـ ريتشاردسون.
جـ . معامل الإستقرار.
د . الإختبار وإعادة الإختبار.

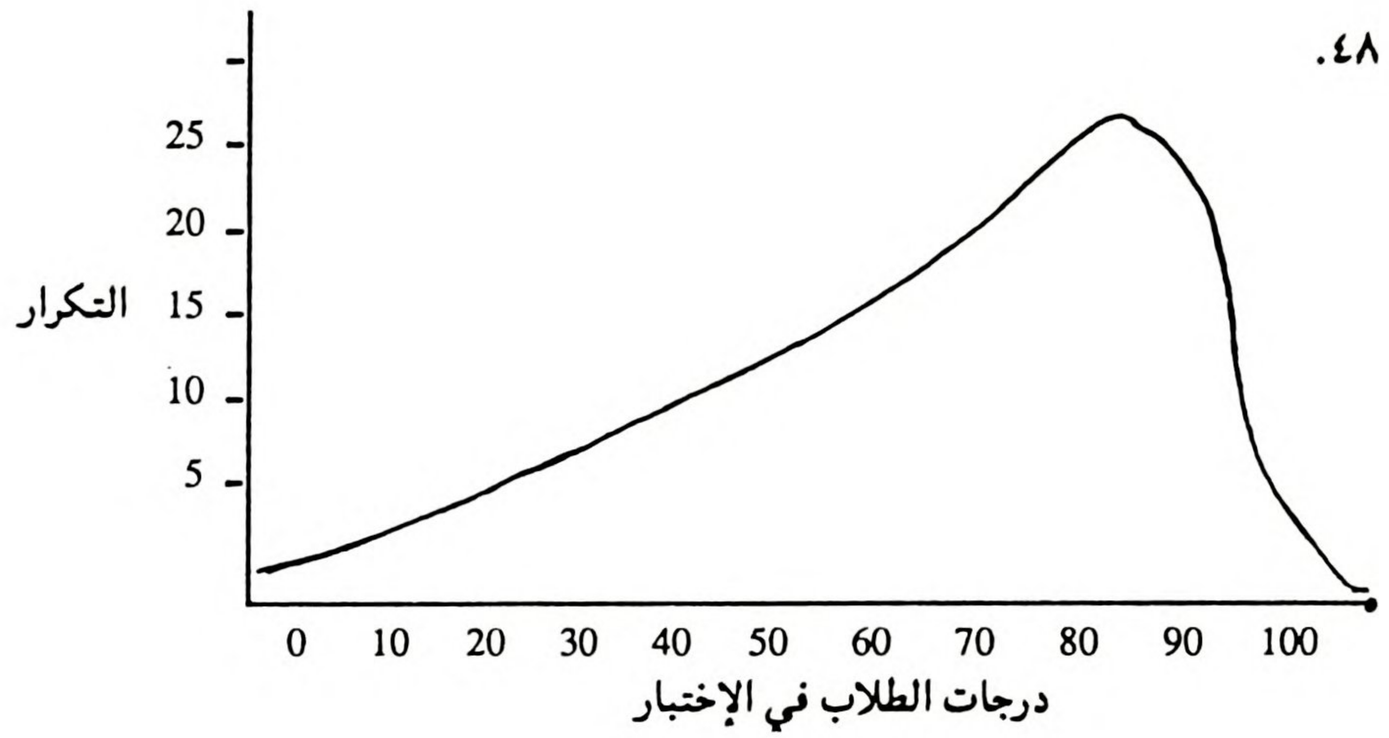
— ٤٤. أحد عيوب الطريقة الجماعيّة المرجع هو أنّ
أ . النسبة المئويه للدرجات لا يمكن حسابها.
ب . رتبة الطالب في الصف لا يمكن تحديدها.
جـ . تقدير الدرجات يكون على شكل تعبيرات (مصطلحات) نسبيّه.
د . قياس الأداء يتم بناءً على معيار ثابت.

— ٤٥. أي من العبارات الآتيه هي الأفضل توضيحاً لتفسير معياري المرجع
أ . حصلت ندى على أعلى درجة في الصف.
ب . أجرت أمل تجربتها في خمس دقائق.
جـ . درجة خالد في الإختبار كانت متوسّطة بالنسبة لصفه.
د . استطاع ابراهيم حل اقل عدد من المسائل الحسابيّه.

ـــ ٤٦. أحد الأعتبارات الهامّه في بناء (إعداد) الأختبار الصفي هو
أ . بناء (إعداد) اسئله تتطلب من الطلاب عمل تمييزات دقيقه.
ب . تطوير نموذجين متطابقين من نفس الإختبار.
جـ . مطابقة الأسئله بالأهداف التعليميّه.
د . إختبار في محتوى لم يتم تدريسه بالصف لمنح الطلاب الفرصه لإظهار تعلّمهم اللاصفّي (خارج الصف).

ـــ ٤٧. أي من إحصائيات السؤال الآتيه يجعل المدرّس أكثر قلقاً؟
أ . مستوى صعوبة السؤال = ٨٠ و.
ب . مستوى صعوبة السؤال = ٥٠و .
جـ . مستـوى تمـييز السؤال = صفر
د . مستـوى تمـييز السؤال = -٥٠و .

ـــ ٤٨.



يشير الرسم البياني أعلاه إلى أنّ الإختبار
أ . سهل جداً.
ب . صعب جداً.
جـ . متوسط الصعوبه.
د . متوسط السهوله.

ـــ ٤٩. التقدير الموضوعي لدرجات الأختبار في أسئلة الأختيار من متعدّد يُعزّزُ
أ . الثبات.
ب . الصدق.
جـ . الصدق ولكن ليس الإستقرار.
د . الثبات والصدق.

ـــ ٥٠. اذا كان ثبات الإختبار يساوي واحد (١)، فإنّ الخطأ المعياري للقياس يكون
أ . صفراً
ب . ١٠ و .
جـ . ٥٠ و.
د . ١

ـــ ٥١. الدرجه التي تقع تحتها ٥٠٪ من الدرجات تُسمّى
أ . الوسيط.
ب . المتوسط.
جـ . المنوال.
د . الإنحراف المعياري.

ـــ ٥٢. أي من الآتي يُعتبر مقياساً لحساب تباين الدرجات؟
أ . الوسط.
ب . الوسيط.
جـ . المنوال.
د . الإنحراف المعياري.

ـــ ٥٣. طبقاً لتصنيف بلوم، في أي مستوى يمكن تصنيف العبارة الآتيه:
«قال المدرس لأحد تلاميذة: إنّ قطعتك في الإنشاء ذات نوعيّةٍ رديئه»؟
أ . التطبيق.
ب . التحليل.
جـ . التركيب.
د . التقويم (التقييم).

ـــ ٥٤. من وجهة نظر القياس التربوي، فإنّ استخدام الإختبارات الصفيّه التي تتكوّن بالكامل من أسئله مقاليه يعتبر غير مرغوب لأن
أ . التصحيح يتطلّب وقتاً طويلاً.
ب . من الصعب جداً بناء (إعداد) الأسئله.
جـ . عَيّنة المحتوى تكون محدوده.
د . بناء (إعداد) نماذج الإجابه يُعتبرُ مضيعةً للوقت.

ـــ ٥٥. أي من الآتي لا يعتبر مرتبطاً بشكل عام بالإختبار المعياري المرجع
أ . مستوى أداء.
ب . مجال محتوى الماده مُحدّد تحديداً جيّداً.
جـ . إختيار أسئلة اختبار طبقاً لمستوى صعوبتها.
د . إستخدام أسئلة الإختبار التي تعكس الأهداف التعليميّة بشكل كبير.

ـــ ٥٦. مجال محتوى الماده المُعَرّفُ تعريفاً رديئاً تنتج عنه أسئله
أ . صعبه جداً.
ب . غير متشابهه.
جـ . غامضه بالنسبة للطالب.
د . غير ثابته.

ـــ ٥٧. عندما نستخدم كلمة موضوعي لتسمية الأختبارات التربويّه، فإنّ هذه الكلمه أكثر ملائمة لوصف.
أ . إحدى خصائص اسلوب التصحيح.
ب . مستوى التقنين في الأختبار.
جـ . ضيق محتوى السؤال.
د . ما إذا كان الإختبار جماعي المرجع أو معياري المرجع.

ـــ ٥٨. أي من الآتي يشير إلى أعلى مستوى للتعلّم بالنسبة للطالب؟
أ . يطبّق مبدأ.
ب . يعطي تعريفاً دقيقاً لمبدأ معيّن.
جـ . يحدّد مثالاً مستخدماً ذلك المبدأ.
د . يعيد صياغة مبدأ ما بأسلوبه هو.

ـــ ٥٩. أيٍ من الأهداف التربويّة الآتيه لا يُعبّرُ عن الأداء ؟
أ . يُصمّم.
ب . يُعدّ.
جـ . يَفهَمُ.
د . يَكتُبُ.

ـــ ٦٠. أحد جوانب القوّه في الإختبارات المعيارية المرجع هو أننا نستطيع
أ . تعميم النتائج على مجموعة أكبر من الأسئله.
ب . مقارنة الطالب بمستوى نسبيٍ للأداء.
جـ . بناء (وضع) معايير موضوعيه للأداء.
د . قياس سَمه (خاصيّه) معيّنه.

ـــ ٦١. تنشأ المشكلات بشكل كبير عند محاولة تطوير مقاييس للأهداف القصوى (البعيده) لأن
أ . معايير الجماعه التي يمكن بواسطتها مقارنة النتائج تكون غير متوافرة.
ب . طرائق (أساليب) القياس لا تعطي وزناً لكل الأهداف.
جـ . مثل تلك الأهداف تتعلق بسلوك لا يمكن مشاهدته من خلال ظروف الصف.
د . المدرسين لا يحبون أن يبتعدوا عن الأساليب التقليديّه للإختبارات.

٦٢. أي من النواتج التعليميّه الآتيه يمكن تصنيفه في المجال الوجداني
أ . تطوير مهارات الكتابه.
ب . تعزيز الميل للموسيقى.
جـ . تحسين الفهم في القراءه.
د . فهم تصنيف الميول.

٦٣. ما المشكلة الكبرى في سؤال الإكمال الآتي بمادة التاريخ للصف الأول الإعدادي: «غزا البرتغاليّون البحرين في ..... »
أ . سهل جداً.
ب . غير واضح.
جـ . به محدّدات للإجابه.
د . يقيس معرفة الحقائق.

٦٤. عند بناء (إعداد) أسئلة الإختيار من متعدّد، فإنّه من الأفضل
أ . جعل كل الأجابات بنفس الطول.
ب . وضع الفكرة الرئيسية للسؤال في الإجابات.
جـ . إستخدام أجابات مثل: أ و ب وليس جـ.
د . أخذ الكلمات الرئيسيه الموجودة في صلب السؤال وإعادة كتابتها أيضاً في الإجابات.

٦٥. في أسئلة الإكمال و الإجابه القصيره يجب أن تُوضَع الفراغات
أ . عند أو بقرب بداية السؤال.
ب . بين بداية ووسط السؤال.
جـ . بقرب ووسط السؤال بقدر الإمكان.
د . عند أو بقرب نهاية السؤال.

انتقل الآن من فضلك ألى أسئلة الإستبانه في الصفحات التاليه

الإســــتبانه

إرشـــــــــادات:

١ - تتكون هذه الإستبانه من ٢٣ سؤالاً.
٢ -من فضلك استخدم الفراغات المعطاه لكتابة إجابتك.
٣ -لاحظ من فضلك أنّ بعض الأسئله تتطلب اختيار رقم واحد فقط من بين الأرقام المقابله للإجابات المعطاه، ثم كتابة هذا الرقم في المكان المُعطى.
٤ - لاحظ من فضلك أنّ بعض الأسئله تتطلب اختيار أكثر من جواب واحد.
٥ - أجب من فضلك عن جميع الأسئله.

•••

١ .العمر؟ ....

٢ .الجنس؟ ....
[١]ذكر
[٢]أنثى.

٣ . أعلى مؤهل تحمله؟ .... **سنة الحصول عليه؟** ....
[١] الدكتوراه.
[٢] الماجستير.
[٣] الدبلوم العالي (ما بعد البكلوريوس / الليسانس).
[٤] البكلوريوس / الليسانس.
[٥] الدبلوم (سنتان بعد الثانويه العامّه)
[٦] شهادة الثانويه العامّه.
[٧] أقل من شهادة الثانويه العامّه.

٤ .فيما يتعلق بأعلى مؤهل أشرت إليه في السؤال رقم (٣)، هل حصلت على ذلك المؤهل من كُليّة تربيه؟ نعم .... لا ....

إذا كانت إجابتك هي (لا)، حدّد من فضلك أعلى مؤهل حصلت عليه من إحدى كليات التربيه. ....

٥ .هل كان تخصصك هو تربيه عندما كنت في السنه النهائيّه بالبكلوريوس / بالليسانس؟ نعم .... **لا** ....

٦ .مادة التخصص؟ ..... .

٧ .المادة التي تدرّسها (للمدرسين)؟ .....
المادة التي تشرف عليها (لإخصائيي المناهج فقط)؟ ..... .

٨ .المرحله التعليميّه التي تدرّسها (للمدرسين) / تشرف عليها (للإخصائيين)؟ ..... .

٩ .عدد سنوات الخبره في التدريس؟ ..... .

١٠.عدد سنوات الخبره كإخصائي مناهج (للإخصائيين فقط)؟ ..... .

١١.ما كمية المقررات التي أكملتها في القياس والتقويم التربوي؟
(يمكنك اختيار أكثر من واحد حسب ما ينطبق عليك).
..... لا شيء.
..... جزء من مقرر آخر.
..... مقرر جامعي واحد.
..... أكثر من مقرر جامعي واحد.
..... مقرر / تدريب أثناء الخدمه.

١٢.ما كمية المقررات التي أكملتها في مبادىء الإحصاء (يمكنك اختيار أكثر من واحد حسب ما ينطبق عليك).
..... لا شيء.
..... جزء من مقرر آخر.
..... مقرر جامعي واحد.
..... أكثر من مقرر جامعي واحد.
..... مقرر / تدريب أثناء الخدمه.

١٣.في أي سنه أنهيت أحدث مقرر في القياس والتقويم التربوي؟ ..... .

١٤.في أي سنه انهيت أحدث مقرر في مبادىء الإحصاء؟ ..... .

١٥.تقريباً ما هى النسبه المئويه من وقتك (داخل وخارج المدرسه) التي تخصصها للإختبارات الصفيّه من حيث الإعداد والتصحيح وتقدير الدرجات وتفسير الدرجات ومراجعة أسئلة الأختبار؟ (أختر إجابه واحده).

..... أقل من ٥٪
..... ٥٪ إلى ١٠٪
..... ١١٪ ألى ٢٠٪
..... أكثر من ٢٠٪

١٦.لأي غرض من الأغراض الآتيه تستخدم الإختبارات الصفيّه التي تعدها بنفسك؟
(يمكنك اختيار أكثر من واحد حسب ما ينطبق عليك).

..... تشخيص جوانب القوه والضعف لدى الطالب.
..... قياس التحصيل الدراسي.
..... قياس إتقان الوحدات التعليميّه بالماده.
..... قياس التقدّم الدراسي.
..... إثارة الدافعيه للتعلم.
..... تقدير ووضع الدرجات.
..... تخطيط التدريس.
..... أغراض أخرى (حدّدها من فضلك) ..... .

١٧.أي نوع من أنواع الأسئله الآتيه تستخدمه لكتابة اختباراتك الصفيّه؟
(يمكنك اختيار أكثر من واحد حسب ما ينطبق عليك).

..... الصواب والخطأ.
..... المطابقه (المزاوجه).
..... الإكمال.
..... ذات الإجابه القصيره.
..... الإختيار من متعدّد.
..... المقاليّه (التي تتطلب إجابتها صفحه أو أكثر).
..... أخرى (حددها من فضلك) .....

١٨. رتب كل نوع من أنواع الأسئله الآتيه من ١ الى ٦ حسب أفضليتّها بالنسبة لك (ملاحظه: ١ = الأكثر تفضيلاً، ٦ = الأقل تفضيلاً).

..... الصواب والخطأ.
..... المطابقه (المزاوجه).
..... الإختيار من متعدّد.
..... الإكمال.
..... ذات الإجابه القصيره.
..... المقاليّه .

١٩. ..... بشكل عام، هل تضع أو تطوّر خطّة إختبار (جدول المواصفات) قبل كتابة أسئلة الإختبار؟
[١] نعم.
[٢] لا.

٢٠. ..... هل تقوم بعمليّة تحليل لأسئلة الإختبارات التي تعدّها أنت؟
[١] نعم.
[٢] لا.

٢١. ..... كيف تُقَيّمُ كفاءتك في القياس التربوي بشكل عام؟
[١] ممتازه.
[٢] جيّده جداً.
[٣]جيّده.
[٤] كافيه.
[٥]ضعيفه.

٢٢. ..... كيف تُقيّمُ كفاءتك في مبادىء بناء (إعداد) الإختبار؟
[١] ممتازه.
[٢] جيّده جداً.
[٣]جيّده.
[٤] كافيه.
[٥]ضعيفه.

٢٣. في أي من الموضوعات الآتيه في القياس التربوي تشعر بأنّك في حاجة إلى تدريب عملي (يمكنك اختيار أكثر من واحد حسب ما ينطبق عليك) ؟
..... تخطيط الإختبارات الصفيّه (مثال: جدول المواصفات).
..... بناء أنواع مختلفه من الأسئله (مثال: الصواب والخطأ).
..... معرفة أي نوع من أنواع الأسئله يجب اختيارها ولماذا، لبناء اختباراتك الصفيّه.
..... إجراء عمليّة تحليل الأسئله لأختباراتك الصفيّه (مثال: مستوى صعوبة السؤال، ومستوى تمييز السؤال).
..... صدق الإختبار (مثال: صدق المحتوى).
..... ثبات الإختبار (مثال: التوافق الداخلي).
..... تفسير درجات الإختبار الصفّي.
..... أساليب الملاحظه.
..... أخرى (حدّدها من فضلك) ..... .

لك خالص شكري وتقديري لتعاونك.

APPENDIX E

MICHIGAN STATE UNIVERSITY

OFFICE OF VICE PRESIDENT FOR RESEARCH
AND DEAN OF THE GRADUATE SCHOOL

EAST LANSING • MICHIGAN • 48824-1046

April 13, 1992

Mr. Rashid H. Aldosary
919-I Cherry Lane
East Lansing, MI 48823

RE: ASSESSING THE COMPETENCY OF TEACHERS, CURRICULUM SPECIALISTS, AND PROSPECTIVE TEACHERS IN EDUCATIONAL MEASUREMENT IN BAHRAIN, IRB #92-167

Dear Mr. Aldosary:

The above project is exempt from full UCRIHS review. The proposed research protocol has been reviewed by a member of the UCRIHS committee. The rights and welfare of human subjects appear to be protected and you have approval to conduct the research.

You are reminded that UCRIHS approval is valid for one calendar year. If you plan to continue this project beyond one year, please make provisions for obtaining appropriate UCRIHS approval one month prior to April 10, 1993.

Any changes in procedures involving human subjects must be reviewed by UCRIHS prior to initiation of the change. UCRIHS must also be notifed promptly of any problems (unexpected side effects, complaints, etc.) involving human subjects during the course of the work.

Thank you for bringing this project to my attention. If I can be of any future help, please do not hesitate to let me know.

Sincerely,

David E. Wright, Ph.D., Chair
University Committee on Research Involving
Human Subjects (UCRIHS)

DEW/pjm

cc: Dr. Irvin J. Lehmann

MICHIGAN STATE UNIVERSITY

DEPARTMENT OF LINGUISTICS AND GERMANIC,
SLAVIC, ASIAN AND AFRICAN LANGUAGES
A-614 WELLS HALL
EAST LANSING, MICHIGAN 48824-1027

Telephone: 517/353-0740
Facsimile: 517/336-2736

April 24, 1992

To Whom it May Concern

This is to certify that I have reviewed the Arabic version of the "Test of Educational Measurement" (MET) written in English by Rashid Aldosary, Ph. D. candidate at the Department of Educational Psychology & Special Education, College of Education, Michigan State University.

I found that the Arabic version is true and correct translation of the original document to the best of my knowledge.

Sincerely,

Dr. Malik Balla
Instructor of Arabic Language.

Department of Linguistics and
Germanic, Slavic, Asian and
African Languages
Michigan State University
East Lansing, MI 48824-1027

Telex 650-277-3148 MCI

Cable: MSUINTPRO ELSG

MICHIGAN STATE UNIVERSITY

COLLEGE OF EDUCATION · DEPARTMENT OF COUNSELING, EDUCATIONAL PSYCHOLOGY AND SPECIAL EDUCATION

EAST LANSING · MICHIGAN · 48824-1034

May 18, 1992

Dr. Abbas Adeeby
Dean of the College of Education
University of Bahrain
Isa Town, BAHRAIN

Dear Dr. Adeeby:

I am writing this letter on behalf of Mr. Rashid Aldosary, a doctoral candidate at Michigan State University, who is working on his doctoral dissertation research.

Rashid is conducting a study on "ASSESSING THE COMPETENCY OF TEACHERS, CURRICULUM SPECIALISTS, AND PROSPECTIVE TEACHERS IN EDUCATIONAL MEASUREMENT IN BAHRAIN". This study will provide data that hopefully, in the coming years, will improve the competence of teachers, prospective teachers, and curriculum specialists in the area of educational measurement. In addition, it is hoped that the results of this study will serve as an important element in bringing about more coordination and integration between the Ministry of Education and the University of Bahrain in planning preservice and inservice measurement training programs for these three groups.

I would be very appreciative if you would permit Rashid to conduct his study on the senior students at your College of Education.

Thank you in advance for your cooperation.

Respectfully,

Irvin J. Lehmann
Professor of Measurement, and
Doctoral Committee Chairperson

MICHIGAN STATE UNIVERSITY

COLLEGE OF EDUCATION · DEPARTMENT OF COUNSELING,
EDUCATIONAL PSYCHOLOGY AND SPECIAL EDUCATION

EAST LANSING · MICHIGAN · 48824-1034

May 18, 1992

Dr. Ebrahim Y. Alabdallah
Director of Curricula
Ministry of Education
Manama, Bahrain

Dear Dr. Alabdallah:

I am writing this letter on behalf of Mr. Rashid Aldosary, a doctoral candidate at Michigan State University, who is working on his doctoral dissertation research.

Rashid is conducting a study on "ASSESSING THE COMPETENCY OF TEACHERS, CURRICULUM SPECIALISTS, AND PROSPECTIVE TEACHERS IN EDUCATIONAL MEASUREMENT IN BAHRAIN". This study will provide data that hopefully, in the coming years, will improve the competence of teachers, prospective teachers, and curriculum specialists in the area of educational measurement. In addition, it is hoped that the results of this study will serve as an important element in bringing about more coordination and integration between the Ministry of Education and the University of Bahrain in planning preservice and inservice measurement training programs for these three groups.

I would be very appreciative if you would permit Rashid to conduct his study on the curriculum specialists and teachers in Bahrain.

Thank you in advance for your cooperation.

Respectfully,

Irvin J. Lehmann
Professor of Measurement, and
Doctoral Committee Chairperson

STATE OF BAHRAIN
MINISTRY OF EDUCATION

دولة البحرين
وزارة التربية والتعليم
إدارة المناهج

بسم الله الرحمن الرحيم

سبتمبر ١٩٩٢م

إسم المدرسة : _____

الأستاذ مدير المدرسة المؤتمر

تحية طيبة واحتراما وبعد ،

يقوم الأستاذ راشد حماد الدوسرى اخصائى القياس والتقويم التربوى بإدارة المناهج بإجراء بحث حول تقييم كفاءة المدرسين فى القياس التربوى ، وذلك لنيل درجة الدكتوراه فى القياس والتقويم التربوى من جامعة ولاية ميتشيغان بالولايات المتحدة الأمريكية .

يرجى ملاحظة ان الأرقام المدونة فى المستطيل أدناه هى الأرقام المتسلسلة للمدرسين بقائمة المدرسين بمدرستكم والذين تم اختيارهم عشوائيا للمشاركة فى الإجابة عن أسئلة أداة البحث (مرفق نسخ من أداة البحث بعدد المدرسين الذين تم اختيارهم).

يرجى التكرم وكتابة رقم المدرس على نسخته من أداة البحث وذلك قبل تسليمه النسخة.

الأرقام المتسلسلة للمدرسين المشاركين فى البحث

وتفضلوا بقبول خالص شكرى وتقديرى لتعاونكم.

مدير إدارة المناهج

ملاحظة هامة :

يرجى التكرم وتسليم جميع النسخ بعد استعادتها من المدرسين إلى مندوب إدارة المناهج ، وذلك بتاريخ ١٩٩٢/٩/١٦م .

سبتمبر ١٩٩٢م

إخصائيو المناهج / المدرسون / الطلبة المعلمون المحترمون.

تحية طيّبة وبعـــــــــــــــــد،

يقوم الأستاذ راشد حمّاد الدوسري إخصائي القياس والتقويم التربوي بإدارة المناهج بإجراء دراسه حول «تقييم كفاءة المدّرسين وإخصائيي المناهج والطلبه المعلّمين (بجامعة البحرين) في القياس التربوي»؛ وذلك لنيل درجة الدكتوراه في القياس والتقويم التربوي وتصميم البحث التربوي من جامعة ولاية ميتشيغان بالولايات المتّحدة الأمريكيه.

اداة الدراسه المرفقه تتكوّن من قسمين. القسم الأول هو «إختبار في القياس التربوي»، والقسم الثاني هو «الإستبانه».

ونظراً لأنكم أنتم فقط الّذين تستطيعون تقديم المعلومات المطلوبه، فإنّ تعاونكم ضروري لنجاح هذه الدراسه.

كل المعلومات التي تدلون بها فى الإختبار والإستبانه سيتم التعامل معها بسريّة تامّه، ولن يطّلع عليها غير الباحث.

يرجى التكرّم والإجابه عن جميع الأسئله في الإختبار والإستبانه.

وتفضّلوا بقبول خالص الشكر والتقدير لتعاونكم.

د. ابراهيم يوسف العبدالله
مدير المناهج

UNIVERSITY OF BAHRAIN
COLLEGE OF EDUCATION
Department of Education

جامعة البحرين
كلية التربية
قسم التربية

التاريخ : ٢٧/سبتمبر/١٩٩٢م .
من : عميد كلية التربية ،
الى : السادة المرشدين الاكاديميين لبرنامج بكالوريوس التربية .
الموضوع: تسهيل مهمه الباحث راشد الدوسرى .

تحية طيبة وبعد

لما كان السيد راشد الدوسرى يدرس بالولايات المتحدة ، ويحتاج الى بعض المعلومات المتعلقه بموضوع رسالته للدكتوراه . وهو بصدد تطبيق أداة بحث على الطلاب الذين تتولون ارشادهم،وحتى يتمكن الباحث من هذا،يرجى مساعدته للاتصال بالطلبة الذين تتولون ارشادهم حتى يحصل منهم على المعلومات اللازمه واننا اذ واثقون من تعاونكم ومساعدتكم للباحث نعبر لكم عن عميق شكرنا وتقديرنا .

مرفق قائمة باسماء الطلبة .

P.O. Box : 32038 - State of Bahrain
Telephone : 440044
Telex : 9552 JAMEA BN
Fax : 442225

ص . ب : ٣٢٠٣٨ ـ دولة البحرين
هاتف : ٤٤٠٠٤٤
تلكس : ٩٥٥٢ جامعة ـ بحرين
فاكس : ٤٤٢٢٢٥

STATE OF BAHRAIN
MINISTRY OF EDUCATION

دولة البحرين
وزارة التربية والتعليم
شعبة القياس والتقويم التربوى

في: ١٩٩٢/١٠/٥م

الرقم: ١/٦٧٦-م/٩٢م

الاستاذ/

الموقر

كلية التربية - جامعة البحرين

تحية أخوية طيبة ملؤها التقدير والإحترام وبعد ،

أود افادتكم باننى أقوم حاليا بإجراء دراسة حول تقييم كفاءة المدرسين وإخصائيى المناهج والطلاب المعلمين (بجامعة البحرين) فى القياس التربوى ، وذلك لنيل درجة الدكتوراه فى القياس والتقويم التربوى من جامعة ولاية ميتشيفان بالولايات المتحدة الأمريكية .

ولقد سبق للأستاذ عميد كلية التربية اشماركم بذلك قبل عدة أيام .

لذا أرجو التكرم وتوزيع النسخ المرفقة من أداة البحث على الطلبة والطالبات المرفقة أسماؤهم وذلك للإستجابة لأسئلة أداة البحث خلال يومين من استلام الطالب لنسخته من أداة البحث وذلك للاهمية .

شاكرا لكم تعاونكم سلفا ، وأملا أن التقى معكم فى حينه .

أخوكم

راشد حماد الدوسرى

إخصائى القياس والتقويم التربوى

هاتف: ٦٣١٠١٠ (المنزل)

٢٤٣٣٧٨ (العمل)

٢٤٤٠١٩ تحويل ١٥٤ (العمل)

STATE OF BAHRAIN
MINISTRY OF EDUCATION

دولة البحرين
وزارة التربية والتعليم
شعبة القياس والتقويم التربوى
إدارة المناهج

في : ١٩٩٢/١٠/٥م

الرقم : ٦٧٦/أ - م/٩٢م

الموقر

الاستاذ /

كلية التربية - جامعة البحرين

تحيـــة أخوية طيبة ملؤها التقدير والإحترام وبعد ،

أود افادتكم بأننى أقوم حاليا بإجراء دراسة حول تقييم كفاءة المدرسين وإخصائيى المناهج والطلاب المعلمين (بجامعة البحرين) فى القياس التربوى ، وذلك لنيل درجة الدكتوراه فى القياس والتقويم التربوى من جامعة ولاية ميتشيغان بالولايات المتحدة الأمريكية .

ولقد سبق للأستاذ عميد كلية التربية اشعاركم بذلك قبل عدة أيام .

لذا أرجو التكرم وتوزيع النسخ المرفقة من أداة البحث على الطلبة والطالبات المرفقة أسماؤهم وذلك للإستجابة لأسئلة أداة البحث خلال يومين من استلام الطالب لنسخته من أداة البحث وذلك للاهمية . يرجى التكرم وايداع النسخ المتبقية مع الطلاب المعنيين مع السيد/ عقيل البوسطه بمكتب عميد كلية التربية .

شاكرا لكم تعاونكم سلفا ، وأملا أن التقى معكم فى حينـــه .

أخوكم

راشد حماد الدوسرى

إخصائى القياس والتقويم التربوى

هاتف : ٦٣١٠١٠ (المنزل)

٢٤٣٣٧٨ (العمل)

٢٤٤٠١٩ تحويل ١٥٤ (العمل)

ملاحظة :
بناءً على اتفاق سبق مع الدكتور حسين بدر، فقد تم وضع نسخة من أسماء الطلاب المرفقة لتتم القبول والتسجيل وذلك للأهمية

STATE OF BAHRAIN
MINISTRY OF EDUCATION

دَولـــة البَحرين
وزارة التربيـــة والتعـــليم
ادارة الماعـــــــــــج
شعبة القياس والتقويـم التربـــــــوى

فى : ١٤/١٠/١٩٩٢م

الأستـاذ / الموقـــــر

كليـة التربيـــة ـ جامعـــة البحريـــــــــــن

أرجو أن تكون قد استلمت رسالـتى المؤرخة فى ٥/١٠/١٩٩٢م ٠ بشأن بحنـــــــى للدكتـوراه ، وتمكنـت من توزيـــع نســخ أداة البحـث على طلابـك الذيـن ترشدهـــــم أكاديميـا وتنطبـق عليهـم شروط البحـث ٠

بعد استلامـك جميـع النســخ من أداة البحـث من طلابـــك بعد استجابتهـــــم لأسئلتهـــا ، أرجو التكرم وايداعهـا مع السيد عقيـل البوسطـــة بمكتـب عميـــــــد كليـة التربيـــة ٠

وتفضلوا بقبول خالص شكرى وتقديـــــرى لتعاونكـــــم ٠

أخـوكـــم
راشـد حمـاد الدوســـــرى
اخصائى القياس والتقويم التربوى

هاتف المنزل : ٦٣١٠١٠
هاتف العمل : ٢٤٣٣٧٨
٢٤٤٠١٩ تحويل ١٥٤

STATE OF BAHRAIN
MINISTRY OF EDUCATION

دولة البحرين
وزارة التربية والتعليم
ادارة الصاعج
شعبة القياس والتقويم التربوى

فى : ١٧/١٠/١٩٩٢م

الأستاذ /

الموقر

كلية التربية - جامعة البحرين

أرجو أن تكون قد استلمت رسالتى المؤرخة فى ٥/١٠/١٩٩٢م ٠ بشأن بحثى للدكتوراه ، وتمكنت من توزيع نسخ أداة البحث على طلابك الذين ترشدهم أكاديميا وتنطبق عليهم شروط البحث ٠ ٠

بعد استلامك جميع النسخ من أداة البحث من طلابك بعد استجابتهم لأسئلتها ، أرجو التكرم وايداعها مع السيد عقيل الوسطة بمكتب عميد كلية التربية ٠

وتفضلوا بقبول خالص شكرى وتقديرى لتعاونكم ٠

أخوكم

راشد حماد الدوسرى

اخصائى القياس والتقويم التربوى

هاتف المنزل : ٦٣١٠١٠

هاتف العمل : ٢٤٣٣٧٨

٢٤٤٠١٩ تحويل ١٥٤

STATE OF BAHRAIN
MINISTRY OF EDUCATION

دولة البحرين
وزارة التربية والتعليم
إدارة المناهج
شعبة القياس والتقويم التربوى

فى : ١٩٩٢/١٠/٢٥م

الأستاذ / الموقر

كلية التربية - جامعة البحرين

تحية أخوية ملؤها التقدير والاحترام وبعد ،

بعد الاتصال ببعض المرشدين الأكاديميين فيما يتعلق بتوزيع نسخ أداة البحث على طلبتهم ، اتضح لى أن هناك صعوبة فى الاتصال بالطلاب المعنيين لكل مرشد أكاديمى .

وبناء عليه ، وكإجراء بديل ، يرجى التكرم وتوزيع نسخ أداة البحث الموجودة لديكم على طلبتكم الذين تدرسونهم حاليا (أثناء المحاضرات) واستلامها منهم بعد الاستجابة لأسئلتها بعد يومين من تسليمها لأولئك الطلبة ، بشرط أن يكون الطالب فى برنامج بكالوريوس التربية ويدرس مساقا واحدا على الأقل فى القياس والتقويم التربوى .

يرجى التكرم وايداع النسخ المستلمة من الطلاب مع السيد عقيل الوسطة بمكتب عميد كلية التربية .

وتفضلوا بقبول خالص شكرى وتقديرى لتعاونكم .

أخوكم

راشد حماد الدوسرى

اخصائى القياس والتقويم التربوى

هاتف :

المنزل : ٦٣١٠١٠

العمل : ٢٤٣٣٧٨

٢٤٤٠١٩ تحويل ١٥٤

STATE OF BAHRAIN
MINISTRY OF EDUCATION

دولـــة البحرين
وزارة التربيـــة والتعـــليم
إدارة المناهـــج
شعبة القياس والتقويم التربـــوى

في : ٤ /١١/١٩٩٢م

الأستاذ/ الموقـــر

تحية أخوية طيبة وبعد ،

أرجو أن تكونوا قد استلمتم معظم نسخ أداة البحث من الطلبة والطالبات الذين وزعتم عليهم تلك النسخ ٠

وأنا اذ أقدر لكم تعاونكم معى وحرصكم على الاستجابة ، أود افادتكم بأن عنصر الزمـــن مهم فى هذه الفترة بالنسبة لى ، وأرجو اخطارى بأى مشكلة تواجهكم فى توزيع النسخ وتـــــرك أى ملاحظـــات حول الموضوع مع السيد عقيل البوسطة٠

أرجو التعاون معى فى جعل آخر موعد لاستلام جميع النسخ من طلبتكم وتسليمها لـــلاخ عقيـــل البوسطة هو يوم الثلاثاء ١٧/١١/١٩٩٢م٠

وتفضلوا بقبـــول خالص شكرى وتقديرى لتعاونكم ٠

أخوكـــم

راشـد حماد الدوســـرى
اخصائى القياس والتقويـم التربـــوى
هاتف المنزل : ٦٣١٠١٠
هاتف العمل : ٢٤٣٣٧٨
٢٤٤٠١٩ تحويل ١٥٤

APPENDIX F

Table 4.14: Item analysis data for the TEST: Teachers.

Item	p	r_{pbis}	item	p	r_{pbis}	Item	p	r_{pbis}
1	.40	-.08	26	.46	-.13	51	.22	.18
2	.39	.32	27	.33	.13	52	.51	.35
3	.49	.26	28	.39	.36	53	.62	.16
4	.30	-.02	29	.66	.28	54	.28	.05
5	.57	.04	30	.11	.10	55	.35	.07
6	.43	.22	31	.34	-.06	56	.05	-.18
7	.28	.20	32	.44	.12	57	.30	.19
8	.58	.06	33	.44	.43	58	.20	.09
9	.39	.42	34	.56	.33	59	.68	.37
10	.44	.10	35	.73	.27	60	.22	-.23
11	.55	.05	36	.10	-.04	61	.39	.29
12	.38	.20	37	.12	-.01	62	.57	.33
13	.48	.36	38	.66	.24	63	.56	.24
14	.36	.11	39	.37	.39	64	.28	.27
15	.58	.32	40	.48	.21	65	.56	.24
16	.20	.12	41	.41	.23			
17	.54	.19	42	.49	.13			
18	.56	.34	43	.16	.15			
19	.28	.07	44	.29	.22			
20	.32	.33	45	.25	-.01			
21	.49	.13	46	.79	.24			
22	.25	-.03	47	.23	.09			
23	.28	.04	48	.31	.18			
24	.57	.17	49	.08	.05			
25	.29	.09	50	.45	.31			

p=Item difficulty index, r_{obis}=Point biserial correlation coefficient.

Table 4.15: Item analysis data for the TEST: Curriculum specialists.

Item	p	r_{pbis}	Item	p	r_{pbis}	Item	p	r_{pbis}
1	.64	.35	25	.32	.13	49	.37	.44
2	.61	.28	26	.58	.15	50	.62	.38
3	.74	.33	27	.70	.55	51	.48	.23
4	.47	.28	28	.74	.23	52	.78	.48
5	.65	.15	29	.88	.03	53	.76	.31
6	.69	.33	30	.06	-.16	54	.30	.36
7	.45	.23	31	.43	.16	55	.50	-.13
8	.49	-.24	32	.50	.16	56	.00	.00
9	.82	.47	33	.82	.44	57	.53	.41
10	.60	.23	34	.71	-.17	58	.36	-.04
11	.40	-.08	35	.90	.32	59	.94	.26
12	.53	.07	36	.15	-.29	60	.09	-.14
13	.72	.35	37	.29	.11	61	.62	.37
14	.45	.50	38	.84	.42	62	.85	.25
15	.76	.52	39	.68	.58	63	.76	.58
16	.40	.36	40	.67	.19	64	.67	.16
17	.74	.37	41	.79	.41	65	.74	.18
18	.70	.17	42	.63	-.09			
19	.41	.19	43	.29	.34			
20	.71	.54	44	.64	.40			
21	.74	.28	45	.52	.46			
22	.47	.11	46	.94	-.04			
23	.16	.33	47	.28	.19			
24	.62	.07	48	.69	.28			

p=Item difficulty index, r_{pbis}=Point biserial correlation coefficient.

Table 4.16: Item analysis data for the TEST: Prospective teachers.

Item	p	r_{pbis}	Item	p	r_{pbis}	Item	p	r_{pbis}
1	.40	-.05	26	.56	.01	51	.33	.34
2	.68	.18	27	.53	.19	52	.62	.09
3	.77	.23	28	.62	.37	53	.80	.36
4	.30	.03	29	.64	.19	54	.36	.26
5	.49	-.12	30	.15	-.01	55	.31	-.01
6	.64	.33	31	.26	-.19	56	.03	-.32
7	.49	.44	32	.53	.03	57	.51	.07
8	.48	-.11	33	.71	.64	58	.26	-.13
9	.70	.21	34	.74	.24	59	.90	.44
10	.42	.04	35	.77	.34	60	.15	.02
11	.70	.19	36	.10	-.23	61	.44	-.05
12	.58	.04	37	.18	.03	62	.81	.41
13	.70	.23	38	.81	.30	63	.80	.19
14	.28	.12	39	.66	.40	64	.57	.27
15	.62	.26	40	.49	.19	65	.55	.25
16	.26	.14	41	.68	.38			
17	.59	.26	42	.60	.09			
18	.75	.37	43	.68	.35			
19	.42	.40	44	.48	.46			
20	.63	.47	45	.28	.11			
21	.70	.33	46	.71	.36			
22	.47	.37	47	.26	.29			
23	.22	.05	48	.44	.41			
24	.39	.09	49	.13	-.02			
25	27	.04	50	.61	36			

p=Item difficulty index, r_{pbis}=Point biserial correlation coefficient.

BIBLIOGRAPHY

BIBLIOGRAPHY

Airasian, P.W. (1990, April). Classroom assessment at the start of school: How teachers size up their pupils. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago.

Airasian, P.W. (1991). Perspectives on measurement instruction. Educational Measurement: Issues and Practice, 10(1), 13-1 6.

Alabdallah, E.Y. (1989). Educational guidance methods development and field follow-up (in Arabic). Unpublished report. Bahrain: Ministry of Education.

Alabdallah, E.Y. (1990). A report on educational guidance and field follow-up for teachers (in Arabic). Unpublished report. Bahrain: Ministry of Education.

American Federation of Teachers, National Council on Measurement in Education, and National Education Association (1991). Standards for teacher competence in educational assessment of students. Educational Measurement: Issues and Practice, 10(1), 30-32.

Boothroyd, R.A. (1990). Variables related to the characteristics and quality of classroom tests: An exploratory study with seventh and eighth grade science and mathematics teachers. Doctoral Dissertation. Albany, NY: State University of New York.

Boothroyd, R.A., McMorris, R. F., & Pruzek, R.M. (1992, April). What do teachers know about measurement and how did they find out? Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco.

Carter, K. (1984). Do teachers understand principles of writing tests? Journal of Teacher Education, 35, 57-60.

Ebel, R. L. (1961). Improving the competence of teachers in educational measurement. The Clearing House, 36(2),67-72.

Ebel, R. L. (1962). Measurement and the teacher. Educational Leadership, 20, 20-24, 43.

Farr, R., Griffin, M. (1973). Measurement gaps in teacher education. Journal of Research and Development in Teacher Education, 7, 19-28.

Fleming, M., Chambers, B. (1983). Teacher-made tests: Windows on the classroom. In W.E. Hathaway (Ed.). Testing in schools: New Directions for Testing and Measurement., No. 19, (pp. 29-38). San Francisco: Jossey-Bass.

Goehring, H.J. (1973). Course competencies for undergraduate courses in educational tests and measurement. The Teacher Educator, 9, 11-20.

Goslin, D.A. (1967). Teachers and testing. New York: Russell Sage.

Green, J.A. (1963). Teacher-made tests. New York: Harper and Row.

Green, K. E., Stager, S. F. (1986-1 987). Testing: Coursework, attitudes, and practices. Educational Research Quarterly, 11 (2), 48-55.

Gullickson. A.R. (1982, November). The practice of testing in elementary and secondary schools. (ERIC Document Reproduction Service No. ED 229391).

Gullickson, A. R. (1984, April). Matching teacher training with teacher needs in testing. Paper presented at the annual meeting of the American Educational Research Association, New Orleans. (ERIC Document Reproduction Service no. ED 249254).

Gullickson, A. R. (1984). Teachers perspectives of their instructional use of tests. Journal of Educational Research, 77, 244-248.

Gullickson, A.R. (1986). Teacher education and teacher perceived needs in educational measurement and evaluation. Journal of Educational Measurement, 23(4), 347-354.

Gullickson, A.R., Hopkins, K.D. (1987). The context of educational measurement instruction for preservice teachers: Professor perspective. Educational Measurement: Issues and Practice, 6(3), 12-16.

Harris, W.S. (1973). Agreement among NCME members on selected issues in educational measurement. Journal of Educational Measurement, 10(1), 63-70.

Hills, J.R. (1991). Apathy concerning grading and testing. Phi Delta Kappan, 72(7), 540-545.

Marso, R.N. (1985, October). Testing practices and test item preferences of classroom teachers. Paper presented at the meeting of the Midwest Educational Research Association, Chicago. (ERIC Document Reproduction Service No. ED 268145).

Marso, R. N., Pigge, F. L. (1987, March). A state-wide assessment of the testing evaluation needs and proficiencies of beginning teachers: Implications for staff development. Paper presented at the meeting of the Association for Supervision and Curriculum Development, New Orleans. (ERIC Document Reproduction Service No. ED 283833).

Marso, R.N., Pigge, F. L. (1987, October). Teacher-made tests and testing: Classroom resources, guidelines, and practices. Paper presented at the meeting of the Midwest Educational Research Association, Chicago. (ERIC Document Reproduction Service No. ED 291781).

Marso, R.N., Pigge, F. L. (1989, March). The status of classroom teachers' test construction proficiencies: Assessment by teachers, principals, and supervisors validated by analysis of actual teacher-made tests. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco. (ERIC Document Reproduction Service No. ED 306283).

Marso, R.N., Pigge, F.L. (1991). An analysis of teacher-made tests: Item types, cognitive demands, and item construction errors. Contemporary Educational Psychology, 16, 279-286.

Mayo. S.T. (1964). What measurement experts think teachers ought to know about educational measurement. Journal of Educational Measurement, 1 (1), 79-86.

Mayo, S.T. (1967). Pre-service preparation of teachers in educational measurement. Final Report, project No. OE4-10-01 1. Washington, DC: United States Office of Education and Loyola University. (ERIC Document Reproduction Service No. ED 021784).

Mayo, S.T. (1970). Trends in the teaching of the first course in educational measurement. Chicago: Loyola University. (ERIC Document Reproduction Service No. ED 047007).

McMorris, R.F., Boothroyd, R.A. (1992, April). Tests that teachers build: An analysis of classroom tests in science and mathematics. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco.

Mehrens, W.A. (1987). Validity issues in teacher licensure tests. Journal of Personnel Evaluation in Education, 1, 195-229.

Mehrens, W.A., Lehmann, I.J. (1984). Measurement and evaluation in education and psychology (3rd Edition). Orlando, FL: Holt, Rinehart, & Winston.

Mehrens, W.A., Lehmann, I.J. (1987). Using teacher-made measurement devices. NASSP Bulletin, 71 (496), 36-44.

Mehrens, W.A., Lehmann, I.J. (1991). Measurement and evaluation in education and psychology(4th edition). Orlando, FL: Holt, Rinehart, & Winston.

Merwin, J.C. (1989). Evaluation. In M.C. Reynolds (Ed.). Knowledge Base for the Beginning Teacher (pp. 185-192). New York: Pergamon.

Newman, D.C. (1981).Teacher competency in classroom testing, measurement preparation, and classroom testing practices. Doctoral Dissertation. Atlanta, GA: Georgia State University.

Newman, D.C., Stallings, W. M. (1982). Teacher competency in classroom testing, measurement preparation, and classroom testing practices. Paper presented at the annual meeting of the National Council on Measurement in Education, New York. (ERIC Document Reproduction Service No. ED 220491).

Nitko, A.J. (1989). Designing tests that are integrated with instruction. In R.L. Linn (Ed.). Educational Measurement (pp. 447-474). New York: American Council on Education and Macmillan.

Nitko, A.J. (1991). What are we teaching teachers about assessment and why? Educational Measurement: Issues and Practice, 10(1), 2.

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

Pigge, F. L., Marso, R.N. (1988, November). Supervisors agenda: Identifying and alleviating teachers' test construction errors. Paper presented at the meeting of the Ohio Association for Supervision and Curriculum Development, Columbus. (ERIC Document Reproduction Service No. ED 304450).

Plake, B.S., Impara, J.C., & Fager, J.J. (1992). Assessment competencies of teachers: A national survey. Lincoln, Nebraska: Kellogg Foundation and Buros Institute of Mental Measurements.

Roeder H.H. (1972). Are today's teachers prepared to use tests? Peabody Journal of Teacher Education, 59, 239-240.

Roeder, H.H. (1973). Teacher education curricula: Your final grade is F. Journal of Educational Measurement, 10(1), 141-143.

Rowntree, D. (1981). A dictionary of education. Totowa, NJ: Barnes & Noble.

Rudman, H.C. (1987a). Classroom instruction and tests: What do we really know about the link? NASSP Bulletin, 71(496), 3-22.

Rudman, H.C. (1987b). The future of testing is now. Educational Measurement: Issues and Practice, 6(3), 5-11.

Rudman, H.C. (1989). Integrating testing with teaching. Washington, DC: Office of Educational Research and Improvement. (ERIC Document Reproduction Service No. ED 315432).

Rudman, H.C., Kelly, J.L., Wanous, D.H., Mehrens, W.A., Clark, C.M., & Porter, A.C. (1980). Integrating assessment with instruction. Institute for Research on teaching., Research Series, No. 75. East Lansing: Michigan State University.

Schafer, W. D. (1991). Essential assessment skills in professional education of teachers. Educational Measurement: Issues and Practice, 10(1), 3-6.

Schafer, W. D., Lissitz, R.W. (1987). Measurement training for school personnel: Recommendations and reality. Journal of Teacher Education, 38(3), 57-63.

Stiggins, R.J. (1985). Improving assessment where it means the most: In the classroom. Educational Leadership, 44(2), 69-74.

Stiggins, R.J. (1988). Revitalizing classroom assessment: The highest instructional priority. Phi Delta Kappan, 69(5), 363-368.

Stiggins, R.J. (1991). Relevant classroom assessment training for teachers. Educational Measurement: Issues and Practice, 10(1), 7-12.

Stiggins R.J., Bridgeford, N.J. (1985). The ecology of classroom assessment. Journal of Educational Measurement, 22(4), 271-286.

Stiggins, R.J., Conklin, N.F. (1988). Teacher training in assessment. Washington, DC: Office of Educational Research and Improvement, US Department of Education. (ERIC Document Reproduction Service No. ED 303439).

Stiggins, R.J., Conklin, N.F. (1992). In teachers' hands: Investigating the practices of classroom assessment. New York: State University of New York Press.

Stiggins, R.J., Conklin, N.F., & Bridgeford, N.J. (1986). Classroom assessment: A key to effective education. Educational Measurement: Issues and Practice, 5(2), 5-17.

Stiggins, R.J, Frisbie, D.A., & Griswold, P.A. (1989). Inside high school grading practices: Building a research agenda. Educational Measurement: Issues and Practice, 8(2), 5-14.

Stiggins, R.J., Griswold, P.A., & Frisbie, D.A. (1986). Inside high school grading practices. Program report., Portland, OR: Northwest Regional Educational Laboratory. (ERIC Document Reproduction Service No. ED 2 79713).

Tyler, R.W. (1951). The function of measurement in improving instruction. In E.F. Lindquist (Ed.). Educational Measurement (pp. 47-67). Washington, DC: American Council on Education.

University of Bahrain. (1991 -1992). College of education catalogue. Isa Town, Bahrain: Author.

Ward, J.G. (1980). Teachers and testing: A survey of knowledge and attitudes. Research Report., Washington, DC: American Federation of Teachers.

Ward, J.G. (1982). An overview of the AFT's "Teaching and Testing". In S.B. Anderson, L.V. Coburn (Eds.). Academic Testing and the Consumer: New Directions for Testing and Measurement., No.16 (PP 47-52). San Francisco: Jossey-Bass.

Yamagishi, M. (1991). Testing in Japan. In K.E. Green (Ed.). Educational Testing: Issues and Applications (pp. 169-195). New York: Garland