1996

Assessing Content-Related Validity and Internal-Consistency Reliability of Tests Constructed by Seychellois Teachers

Justin Davis Valentin
Edith Cowan University

Edith Cowan University

Copyright Warning

ASSESSING CONTENT-RELATED VALIDITY AND INTERNAL-CONSISTENCY RELIABILITY OF TESTS CONSTRUCTED BY SEYCHELLOIS TEACHERS.

By

Justin Davis Valentin, Dip Ed.

A Thesis Submitted in Partial Fulfilment of the Requirements for the Award of Bachelor of Education with Honours

Faculty of Education

Edith Cowan University

Date of submission: ..4th July., 96.....

USE OF THESIS

The Use of Thesis statement is not included in this version of the thesis.

Abstract

Teachers are seldom trained to construct tests (Boothroyd, McMorris, & Pruzek, 1992; Wise, Lukin, & Roos, 1991). Yet, the use of teacher-made tests for assessing students is a common occurrence in schools. This study challenges the quality of tests constructed by teachers without measurement and testing training.

A sample of tests (n = 15) constructed by Primary 5A mathematics teachers in the Seychelles was analysed. The teachers who submitted the tests have not completed a course in measurement and testing. However, results of these tests will be used to make important decisions about their students.

The purpose of the study is to ascertain whether tests constructed by teachers without measurement training produce valid and reliable scores.

The findings of this study indicate that the test results have high internal consistency reliability, low content-related validity, and a low percentage of effective items. Hence, recommendations are made to the School of Education in the Seychelles to assist teachers in test construction and to include a measurement course in its pre-service teacher training program. It is recommended that in-service teachers use other forms of assessment instruments because the study shows that tests alone are not adequately measuring the students' performances.

Declaration

I certify that this thesis does not incorporate without acknowledgment, any material previously submitted for a degree or a diploma in any institution of higher education; and that to the best of my knowledge and belief it does not contain any material previously published or written by another person except where due reference is made in the text.

Signature...

Date.....4th July, 96.....

Acknowledgment

This work would not have been completed successfully without the assistance of the some people and organisations. I owe many thanks to all of them.

I would like to thank my supervisor, Dr. John Godfrey for all his help during this thesis preparation. I extend my thanks to Dr. Ken Knibb for all his advice.

I would like to thank all the teachers and their students for providing me with the relevant materials I needed for the research. I thank my girlfriend Monica Zelime for assisting me in filling the item matrix tables for the tests analysis, and my friend Daniel Addrienne for assisting me in assessing the content-related validity evidence of the tests.

I would like to express my appreciations and thanks to the Ministry of Education and Culture in Seychelles for its financial support. I extend my thanks to Mrs. Dolores Azemia, Mrs. Gopal, Mrs. Rose-Helen Sinon, and Miss Aglae from the Ministry of Education and Culture.

I would like to thank AusAid for sponsoring my course.

Dedication

I dedicate this thesis to my girlfriend Monica. It is because of her continuous exam pressures and complaints that I have come up with this study. I also dedicate this thesis to anyone else who believes that it was high time to "get ladan* ".

* *"get ladan" is a very popular Creole expression used when an irritating matter needs an immediate investigation.*

Table of Contents

List of Figures

List of Tables

CHAPTER 1

Introduction

Overview of the Thesis

This is a study carried out to ascertain whether tests constructed by teachers without measurement and testing training produce valid and reliable results. The main subjects were mathematics teachers from Primary 5A classes of the Seychelles. The teachers submitted a copy of the maths test they administered to their students as an end of year test. They also submitted the students' marked answer papers and the objectives the tests were assessing. The tests were analysed to determine their degree of content-related validity evidence, internal consistency and equivalent form reliability of their scores, and the effectiveness of their items.

This chapter outlines the background of the study, its significance, purpose, and the research questions. It also includes a section which defines the measurement terminologies which are used throughout the thesis.

Chapter 2 provides a literature review. The main headings of the literature review are: the functions of teacher-made tests in classrooms, the major characteristics of a well-designed test, the nature and quality of teacher-made tests, the teachers' testing practices, experts' views of what teachers should know about measurement, and some suggested methods of evaluating the internal consistency reliability and content-related validity evidence of test results.

In Chapter 3 the theoretical framework of the study is outlined. The methods and procedures used to carry out the research are described in Chapter 4. Chapter 5 reports the results of the study. Discussions of the results are included in Chapter 6 together with the limitations of the study and recommendations made from the results obtained.

Background of the Study

In primary and lower secondary schools in the Seychelles, teacher-made tests are the main instruments used by teachers to gather information about students' performances. For some teachers, a teacher-made test is their only assessing instrument.

In all schools in the Seychelles, teacher-made tests are administered for four main purposes. At the beginning of the year or term teachers administer teacher-made tests to organise students in different class groups. Prior to the beginning of a topic, they are administered to measure the students' prerequisite knowledge and at the end of the topic they are administered to measure the students' achievement. Teacher-made tests are also administered at the end of each term to report students' progress to parents and to stream students. Teacher-made tests play a central role in the Seychelles school system in the evaluation of students' learning.

Students take non teacher-made tests at the end of the third term of Primary 4, Primary 6 and Secondary 4. The tests they take at these levels are constructed by the National Curriculum Department and the results are used in evaluating the effectiveness of the programs set for each of these three stages.

Teacher-made tests help define the future of the students. When students commence lower primary, they are grouped randomly in different classes. They are regularly given teacher-made tests. Their performances on these tests provide feedback about their learning progress and their abilities relative to other students at the year level. Towards the end of lower primary, the teacher-made tests determine how they will be organised into different year groups; Class A, Class B, Class C and Class D in the following year level.

The highest achievers are placed in Class A while lowest achievers are placed in Class D.

At other year levels, teacher-made tests are used for similar functions. In addition, the teacher-made tests administered at the end of each term determine whether the students will be promoted or relegated to a different class group in the following term. Moreover, in the lower secondary schools the results of these tests help to identify students to be placed in the vocational or the academic channel of the upper secondary. The students in the vocational channel are prepared to enter the world of work after completing their secondary studies while those in the academic channel are prepared to follow further studies.

The Problem

Most teachers currently teaching in both the primary or lower secondary schools in the Seychelles are locally trained teachers. The teacher training programs they followed did not include a course in measurement and testing nor were they formally trained to construct tests. Even in the in-service teacher education programs they are not given a test construction training segment. There is no verification whatsoever to determine whether they are constructing tests that adequately measure the students' academic achievement. Yet, the results of these tests are used in making decisions about students' learning and their future.

A recommendation was made to the Ministry of Education in the Seychelles in 1990 to improve the quality of classroom assessment (University of Cambridge Local Examination Syndicate [UCLS], 1990). A seminar on principles and techniques of assessment was conducted by consultants from the University of Cambridge Local Examination Syndicate. Following the seminar the consultants made recommendation to the Ministry to create an assessment unit within the Educational Planning and

Development Division whose aim would be designing, establishing, and monitoring an assessment system in the school. Unfortunately, the recommendation has not been implemented and the reasons for the omission cannot be determined. It is believed that the Ministry officials or school authorities do not see a need for training teachers in test construction. Or they believe that the training given in the teacher training institution adequately prepares teachers to construct and use tests.

Purpose of the Study

The purpose of the study is to analyse a sample of teacher-made mathematics tests constructed by locally trained Seychellois teachers to determine their degree of content-related validity evidence, and the internal consistency reliability of their results, and thereby ascertain whether these tests adequately and fairly measure the student learning.

Significance of the Study

Since no previous studies of analysing teacher-made tests has been conducted in the Seychelles, the research will provide information about the quality of tests that are used in the primary and the lower secondary schools. The results of the study will also provide evidence as to whether recommendations to assist in-service teachers about test construction, and propositions for a course in measurement and testing in the teacher training program can be made.

The Research Questions

Many teacher training programs continually exclude a course in measurement and testing as a requirement for certification although experts in educational

measurement (Gay, 1991; Sax, 1989; Stiggins, 1994) claimed that teachers should be given such training in order to construct effective tests. The major research questions that the study intends to answer are:

1. Do tests constructed by teachers without training in measurement and testing produce valid and reliable results?

2. Can results of these tests be used in making decisions about the students' learning?

The subsidiary questions in this study relate to the content-related validity evidence, internal consistency and equivalent form reliability, and effectiveness of the items of the tests. The subsidiary questions are:

1. What is the degree of content-related validity of the tests?

2. What is the internal consistency reliability of the results of these tests?

3. What is the degree of equivalent form reliability between Paper 1 and Paper 2 of all the two-paper tests?

4. What does item analysis indicate about the effectiveness of items used in these tests?

5. What are the teachers' strengths and weaknesses in test construction evident from the sample of tests analysed?

Definitions of Terminologies Used

This section presents the definitions of terminologies used throughout the thesis. The terms defined are mainly those used in measurement and testing contexts.

Assessments are the full range of procedures used by teachers to gain information concerning the student learning progress (Linn & Gronlund, 1995, p. 5).

Content-related validity is the degree to which a test measures an intended content area (Linn & Gronlund, 1995, p. 51). It also explains the extent to which the

test requires demonstration by the students of the achievements which constitute the objectives of instruction in this area (Ebel, 1983, cited in Hopkins, Stanley, & Hopkins, 1990).

Discriminating power is a value which indicates the degree to which an item distinguishes between the students who mastered and those who did not master the objectives of a test (Gay, 1991, p. 253).

Distracting power indicates the extent to which an option of multiple choice questions attracts the examinees (Gay, 1991, p. 255).

Evaluation is the systematic process of collecting and analysing data for making judgements (Gay, 1991, p. 6).

Factor analysis refers to a variety of statistical techniques whose common objective is to represent a set of variables in terms of a smaller number of hypothetical variables (Kim & Mueller, 1978, p. 9).

Internal consistency reliability refers to the correlation or consistency among the items on single tests (Worthen, Borg, & White, 1993, p. 144). Thus students who do best on one quality that is being scored tend to be the students who do best on the other qualities (Oosterhof, 1994).

Item analysis is a name given to a variety of statistical techniques designed to assess individual items on a test after the test has been given to a group of students (Oosterhof, 1994, p. 195).

Item difficulty indicates the proportion of students on a test who responded correctly to a particular item (Gay, 1991, p. 252).

Measurement and testing course is an area of study which examines the theory, construction and use of tests, and other evaluation instruments (Green, 1989).

Multidimensional scaling is a set of mathematical techniques that enable a researcher to uncover the hidden structure (for example, examinees' responses) of data bases (Kruskal & Wish, 1978, p. 5).

Reliability is the degree to which a test consistently measures whatever it measures (Gay, 1991, p. 166).

Teacher-made tests are tests that have been constructed (written or assembled) by the classroom teacher to assess students (Worthen, et al., 1993, p. 78).

Untrained teachers in this research refers to teachers who have not formally completed a course in measurement and testing.

Validity is the degree to which a test measures what it is supposed to measure (Gay, 1991, p. 157). Validity is viewed as a unitary concept based on various kinds of evidence namely, content-related, criterion-related, construct-related, and consequences validity evidence (Linn & Gronlund, 1995, p. 50).

CHAPTER 2

Literature Review

Overview of the Chapter

This literature review focuses on five issues: the functions of teacher-made tests in classrooms; the major characteristics of a well-designed test; the teachers' tests and testing practices; views of some measurement experts on what teachers ought to know about measurement and testing; and some suggested methods of evaluating the internal consistency reliability and content-related validity evidence of test results.

As no research has focused on measurement and testing in the Seychelles, the review treats the above issues globally. Most of the studies presented in this chapter were carried out in the United States whereby researchers investigated the testing practices in the various states.

The Functions of Teacher-Made Tests in Classroom

Teacher-made tests for the purpose of judging students' learning are common occurrences in the classroom (Fleming & Chambers, 1983). Boothroyd, McMorris and Pruzek (1992) reported that between 90% to 95% of teachers regularly construct their own tests to assess the students' competency.

According to Hopkins, Stanley, and Hopkins (1990), teacher-made tests serve two major functions- instructional and administrative functions. The instructional functions include providing the instructor and students a means of feedback, a way of motivating the students to learn, and consequently facilitating the learning process. The administrative functions provide a mechanism of quality control by monitoring

the achievement of learning. They also facilitate better classification and placement decisions.

Gullickson (1984) reported that teachers believed that the tests they constructed increased the students' effort, influenced the students' self concept, improved interaction among students and consequently the learning environment, provided good learning experiences for students, and accurately revealed students' progress. The students interviewed reported that teacher-made tests assisted them with their learning.

Teacher-made tests are effective if the tests are well designed. Well-designed teacher-made tests have many advantages over other forms of tests. Since they match well with the instructional objectives, their results can be used to provide feedback to both students and teachers regarding the learning progress (Linn, 1983). Thus they are useful for formative evaluation and can provide a basis for reporting students' achievement to parents (Satterly, 1981). Research has also shown that a test skilfully constructed by teachers can be as precise as standardised tests (Hopkins, et al., 1990). When content is clearly specified to students and when tests are carefully developed from these specifications, students' learning is rapid and dramatically positive (Roid & Haladyna, 1982). Roid and Haladyna (1982) argued that tests based on systematic item development will be accurate in providing feedback to students.

Characteristics of a Well-Designed Test

Although it is difficult to evaluate a test as good or bad (Griffin & Nix, 1991; Worthen, Borg, & White, 1993) there are some agreed features that any test should possess. For instance, Tuckman (1988) argued that a test needs to be representative of students' proficiency and the results should have credibility. Carey (1994) added that good tests should be linked with the content of the course. This ensures that what

was taught is tested and what is tested was taught. If not, as Nitko (cited in Worthen, et al., 1993, p. 238) noted,

> The failure to appropriately link testing and teaching will often lead to situations where: (1) students' motivation for learning is reduced; (2) incorrect information is given about students' learning progress and difficulties; (3) critical decisions about the promotion may be made unfairly; and (4) incorrect decisions may be reached about instructional effectiveness.

Since teacher-made tests are used for evaluating students' learning, they should have a high degree of content validity evidence. Worthen, et al. (1993) stated, "Ideally a test should sample all important aspects of the content domain. No important parts of the domain should be under represented or excluded. Similarly, no aspect of the domain should be over represented" (pp. 180-181). Tuckman (1988) believed that content validity is the primary evidence of validity that teachers can provide when constructing a test. A test should also have face validity. Face validity refers to the degree to which a test appears to measure what it purports to measure. Hopkins, et al. (1990, p. 79) pointed out that "it is important for a test to have face validity; otherwise, students may feel that they are being unfairly assessed."

Reliability is another essential characteristic of the results of an effective test. Reliability refers to the degree to which a test consistently measures whatever it measures (Airasian, 1994; Gay, 1991). In this research the reliability in question is mainly the internal consistency reliability. A test has internal consistency reliability if everything that contributes to the score is related. Thus students who perform best on one quality that is being scored tend to be the students who perform best on the other qualities (Oosterhof, 1994).

The Teachers' Tests and Testing Practices

Fleming and Chambers (1983) reported that little is known about the nature and quality of teacher-made tests. According to Stiggins and Bridgeford (1985), and Stiggins (1991) a possible reason for a lack of knowledge about the nature and quality of teacher-made tests is that research on classroom assessment has tended to concentrate on standardised tests and has paid minimal attention to teacher-developed assessments. Also, much of what is known about teachers' tests and testing practices has been obtained through studies using teacher self-reporting data-gathering procedures. These self reporting data gathering studies provide a valuable but limited understanding of teachers' actual tests (Marso & Pigge, 1992).

Marso and Pigge (1992) claimed that the nature and quality of teacher-made tests are unknown. Few studies have been done whereby researchers directly assessed teacher-constructed tests with the exceptions of Green and Stagers (1986), Marso and Pigge (1992), McMorris and Boothroyd (1992), and Stagers and Green (1984).

Despite the scarcity of knowledge about the nature and quality of teacher-made tests, there are some consistencies in the few studies available. For instance, Fleming and Chambers (1983) reported that the most frequently used item format in the United States is short answer, followed by matching, multiple choice, true or false, with essay questions being the least used format. This is consistent with the studies of Marso and Pigge (1988), McMorris and Boothroyd (1992), and Oescher and Kirby (1990).

These studies also showed that most of the items function at a low cognitive level. However, the study of Fleming and Chambers (1983) reported that some items on mathematics tests were functioning at a higher cognitive level but on the whole the tests were poor. Most of these tests contained writing flaws and item writing rule

violations. The results of a study by Kirby and Oescher (1987) showed that poor quality test items are due to the fact that teachers do not put enough effort into test construction. Teachers seldom prepare plans or tables of specifications nor do they validate their tests (Marso & Pigge, 1988, 1989).

Many teachers do not feel confident about their ability to write good tests (Carter, 1984; Stiggins & Bridgeford, 1985). Some teachers believe that the training they received in testing was somewhat below the training they received in other areas of teacher education (Gullickson, 1984). The literature shows that in general the testing skills of teachers in the United States are inadequate for their testing practice (Kirby & Oescher, 1987; McMorris & Boothroyd, 1992).

The teachers' beliefs about their testing practices have not been consistent over research findings. Green and Stager (1986) reported that teachers felt confident about their testing practices and did not see the need of further training in testing. Kubizyn and Borich (1993) reported that teachers claimed they do not need a course in testing or measurement because testing is merely supplemental to instructional process.

Rather surprisingly, Newman and Stallings (1982) found that teachers who had completed a course in measurement and testing rarely constructed good tests. Stiggins (1991) believed that there is a mismatch between what is given in teacher training programs and what teachers need in real classroom practice. Thorndike, Cunningham, Thorndike, and Hagen (1991) supported this argument. They added that the agreed-upon methodology for constructing and the methodology for conducting analyses of test results are not easily understood. Moreover, they argued that even if completely understood these methodologies are time consuming.

Commenting about the teachers' testing and assessment skills, Stiggins (1994) concluded that teachers are not assessment literate. He defined assessment literate

teachers as "those [teachers] who understand the basic principle of sound assessment and can point to others when their assessment fails to measure up" (p. 6). He argued that a carefully planned course in measurement and testing will certainly increase the teachers' testing skills and will eventually make them better assessors.

Requirements of a Measurement Course

Farr and Griffin (1972) believed that teachers should realise how measurement can be used to help them improve their instructional planning and decision making. In addition Gullickson and Ellwein (1985) contended that measurement courses should also equip teachers with skills for post hoc statistical analysis of tests. They argued that such analyses are essential to summarise students' performances, to assure the validity of a test, and to improve the quality of individual items. In advancing their argument they cited Gronlund (1981) who pointed out that item analyses provide data as a basis for efficient class discussion of test results, general improvement of class instruction, and increased skills in test construction.

Gullickson (1986) surveyed a sample of teachers and academics about what they considered should be offered in measurement courses. The teachers and the academics agreed that teachers should be trained about how to plan and construct classroom tests. Since tests have an impact on students, they believed teachers should be capable of producing tests that have few negative implications.

Secondly, they believed that a measurement course should give teachers skills to use non-test evaluation procedures such as observational techniques and ratings. They agreed that teachers should be trained to use assessment results for instructional planning and formative evaluation.

Another area which they believed a measurement course should treat is the use of assessment for summative evaluation. They argued that because grades and other

forms of summative evaluation are used in judgements and decisions concerning students, teachers should understand the different approaches to marking and combining marks so that the summative evaluation validly and fairly reflects the students' accomplishments (Linn, 1990).

The teachers and the academics also believed that a measurement course should provide teachers with knowledge about how to administer and score tests. They also agreed that teachers need to know about general information regarding the selection and use of tests (Frary, Lawrence, & Weber, 1993).

Although there are differences in what experts in measurement believe teachers ought to know about measurement, their views seem to focus on the need for teachers to understand how to link assessment and course content which truly measure their students' performances. As Stiggins (1991) asserted, teachers should be trained to understand the meaning of quality assessment and the importance of designing assessment with a clear vision of the achievement target.

Evaluating Internal Consistency Reliability and Content-Related Validity

Internal Consistency Reliability

There are three methods commonly used for evaluating the internal consistency reliability of tests (Frisbie, 1988). All these methods require only one administration of the tests. The split half method involves splitting the tests into halves. A score on each half is obtained for each examinee and the half scores are correlated to obtain a correlation coefficient. If the results have internal consistency reliability, students who score high marks on one half tend to score high marks on other half. The reliability coefficient of the whole test is finally adjusted using the Spearman-Brown formula, $\underline{r} = 2\underline{r}_{\underline{h}} / (1 + \underline{r}_{\underline{h}})$, where $\underline{r}$ is the reliability of the full-length test and $\underline{r}_{\underline{h}}$ is the actual correlation between the two half-tests (Thorndike, et al.,

1991). The adjustment is necessary because split half method under-estimates the reliability of the test.

The Kuder-Richardson method uses the following formula: $r_{\text{total test}} = [K(SD)^2 - \mu(K - \mu)] / [(SD)^2 (K - 1)]$ where $r_{\text{total test}}$ is the reliability estimate, K is the number of items on the test, μ, the mean of the test scores, and SD, the standard deviation of the scores (Frisbie, 1988). This reliability estimate is equivalent to the average of the split half reliability values for all possible halves of a test. The limitation of this method is that it makes use of a formula which can only be applied to tests in which their items are dichotomously scored (Worthen, et al., 1993).

A third method, Cronbach's coefficient alpha, is widely used nowadays. This method is a generalised form of Kuder-Richardson. This method is used with tests regardless of the scoring or weights of items.

Content-Related Validity

The methods for evaluating content-related validity evidence reported in the literature are classified as empirical methods and subjective methods. Empirical methods analyse the examinees' responses to test items. Factor analysis or multidimensional scaling is used to analyse the inter-item correlation matrix derived from the examinees' responses (Sireci & Geisinger, 1992). The resulting factors or dimensions are compared with the structure of the content domain specified in the blueprint. Although these methods are objective (Sireci & Geisinger, 1992) they are criticised because the degree of relevance of an item to its corresponding content domain is a concept that is independent of the examinees' performances on the item.

The subjective methods use subject matter experts who review test items and rate them according to their degree of appropriateness for measurement of the content domain they purport to measure. For instance, Aiken's validity index (Aiken, 1980),

provides an indication of how different subject matter experts rate the relevance of an item to a particular content domain (see Equation 1). In Equation 1, V is the validity index, i refers to a particular category on the scale (usually 0,1,2.....), c is the number of categories on the scale used to rate the item, n_i is the number of judges who rate an item into the ith category, and N is the total number of judges.

$$V = \frac{\sum_{i=1}^{c-1} i n_i}{N(c-1)} \qquad (1)$$

The Percentage of Items is another method used for obtaining an index (Equation 2) which indicates the proportion of items that assess the objectives of the curriculum (Crocker, Miller, & Franks, 1989). The basic data for computing this index are obtained when one content area expert considers each item and makes a dichotomous decision about its match to a list of objectives. In Equation 2, P is the index, n_j is the number of items matched to any objective by judge j, N is the number of items on the test, and J is the number of judges.

$$P = \frac{\sum_{j=1}^{J} n_j}{NJ} \qquad (2)$$

These two methods and other subjective methods reported in the literature are used mainly on standardised tests for assessing the fit between the tests and the school curriculum. Their limitations are that they assess an item in isolation (Crocker, et al., 1989). They provide indices which indicate the extent to which an item matches with curriculum but do not indicate how well the items of the tests are sampled. Also, the resulting indices are most significant when many subject matter experts are involved in the judging procedures.

Conclusion

A search of literature shows that limited research has been conducted in the area of classroom assessment; thus, it is difficult to fully ascertain the nature and quality of teacher-made tests. Further, the few studies reported in the literature, have focused more on ascertaining the nature of tests in terms of the cognitive level of the items relative to Bloom's taxonomy of educational objectives (Bloom, Madaus, & Hasting, 1981).

In the Seychelles no study has been conducted whereby researchers evaluated the content-related validity or calculated the internal consistency reliability of the teacher-made tests. Thus, the significance of this study is justified.

The literature provides sufficient evidence to support the claim that teacher training programs fail to equip the teachers with good testing skills. As Lindquist (cited in Stiggins, 1994, p. 1) stated,

> If measurement is to continue to play an increasingly important role in education, measurement workers must be much more than technicians. Unless their efforts are directed by sound educational philosophy, unless they accept and welcome a greater share of responsibility for the selection and clarification of educational objectives, unless they show much more concern with what they measure as well as with how they measure it, much of their work will prove futile and ineffective.

In relation to deciding whose responsibility it is to ensure that teachers receive such training, Gay (1991) believed that teacher training institutions, in-service educational programs, and teachers themselves share equal responsibilities.

CHAPTER 3

Theoretical Framework

The Theoretical Assumptions of the Study

The literature indicates several factors that influence the quality of teacher-made tests. According to Sax (1989) the effectiveness of teacher-made tests depends on the teachers' skills and knowledge of test construction. Wesman (1971) believed that teachers must have a thorough mastery of the subject matter being tested. He argued that teachers must not only be acquainted with facts and principles but they must also be fully aware of their implications-- the popular fallacies and the misconceptions in the field. It is suggested that the quality of tests will improve when teachers have knowledge of both the subject matter and test construction.

It is reported in the literature review of this paper that teachers who have had training in test construction do not necessarily construct good tests (Newman & Stallings, 1982). This suggests that there are other factors affecting the test construction. Some of these factors may be the teachers' unwillingness to fully apply themselves in test construction (Kirby & Oescher, 1987; Marso & Pigge, 1988) or the teachers' inability to apply their knowledge (Thorndike, Cunningham, Thorndike, & Hagen, 1991). The literature also reports a mismatch between what teachers received in measurement training and what they actually need in classroom practices (Gullickson, 1984; Stiggins, 1991). These factors are termed, actual practice.

It is assumed that there are other factors affecting the quality of tests. Such factors include the teachers' knowledge of the students for whom the test is intended, teachers' work load, time, and resources. Thus, the combination of teachers' knowledge of subject matter and test construction, the teachers' actual practice, and other related factors result in good quality tests (see Figure 1).

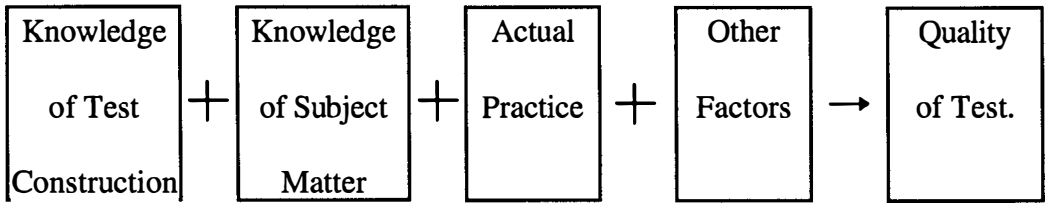

Figure 1. A suggested combination of factors which influence the quality of teacher-made tests.

In this research the quality of tests is measured by content-related validity evidence, internal consistency reliability of the results and effectiveness of the items. The tests were analysed to ascertain the degree to which their items represent the entire body of content to be measured. This provided a degree of content-related validity evidence. For internal consistency reliability, a coefficient alpha value was calculated for the results of each test. In addition equivalent form reliability was calculated for all two-paper tests. Lastly, item analysis was performed on all items of each test.

Interpreting the Quality of Tests

Worthen, Borg, and White (1993) claimed that a test has high content-related validity evidence if it samples all important aspects of the content domain. All parts of the domain should be well represented

Item analysis provides information about the effectiveness of items of the test. Item difficulty is a vital component of item analysis and it is expressed as the proportion of examinees getting an item correct. As the tests analysed were end of term tests constructed to test achievement and to compare students' performances

relative to others, a favourable index between 0.2 and 0.8 exclusive was considered as an acceptable value. An item difficulty index below 0.2 or above 0.8 indicates that the item is too difficult or too easy respectively.

The discriminating index is another vital component of item analysis. A perfect discriminating index is 1.0, indicating that top achievers on the test are getting an item correct and low achievers are getting the item wrong. However, this is not always the case in a test. For teacher-made tests, an index above 0.3 is generally considered sufficient (Oosterhof, 1994). If the index goes below 0.2, then the item is rejected. An item with an index between 0.2 and 0.29 may be revised and used again.

Ebel and Frisbie (1991) pointed out that there are no standards to serve as criteria for determining whether a given reliability coefficient is high enough. However, they argued that some relative standards have evolved over time for evaluating reliability under several circumstances. For instance, teacher-made tests tend to produce scores with a reliability coefficient around 0.70. Most educators regard this value as acceptable. However, experts in educational measurement (Ebel & Frisbie, 1991; Guilford, 1965) believe that the reliability coefficient should be between 0.70 and 0.98 if the scores are to be used in making decisions about individuals and are the only available useful information. If a decision is based on the scores of a group of individuals then the generally acceptable minimum standard is considered to be as low as 0.65 (Ebel & Frisbie, 1991).

Summary

The quality of teacher-made tests is believed to be affected by the following factors: teachers' knowledge of test construction and subject matter; teachers' ability to fully apply their knowledge in actual practice; and other factors which include teachers' knowledge of the students to be tested, time, teachers' work load, and

resources. The quality of tests in this study is determined by assessing the degree of content validity, calculating the degree of internal consistency and equivalent form reliability of the test results, and performing item analysis on items of each test.

CHAPTER 4

Methods and Procedures

This chapter contains a description of the methodology, the subjects chosen, the instrument administered, and the data analysing procedures used in this study.

Methodology

All Primary 5A mathematics teachers across the Seychelles (n = 23) were invited to participate in the study. They were required to submit a copy of the test they administered to their students at the end of the 1995 school year, the objectives their tests were assessing, and the students' marked answer papers.

The tests were directly analysed to determine their degree of content-related validity, internal consistency and equivalent form reliability of their scores. Item analysis was performed on the items of all the tests to ascertain their degree of effectiveness.

Marso and Pigge (1992) noted:

> Much of what is known about teachers' tests and testing practices has been obtained through studies using teachers self-report data-gathering procedures. These self-reported studies provide a valuable but at best a limited understanding of teachers' actual testing knowledge and skills. Very few studies involving direct observations of teachers' testing practices or involving direct analyses of teacher-constructed tests appear in the literature. (pp. 3-4)

This conclusion made by Marso and Pigge (1992) is a major justification of the methodology used for this study.

In order to increase homogeneity in the research, the study was limited to only one year level and one subject area. Maths tests were preferred because maths tests constructed at this level are usually objective.

Description of the Subjects

The teachers who participated in the study (n = 15) received their teacher training in the Seychelles. There were two (13%) male teachers and thirteen (87%) female teachers.

A teaching colleague assisted the researcher in assessing the content-related validity evidence of the test results. The researcher and the teaching colleague have similar experience and qualifications in teaching. Both teach maths and both have completed courses in measurement and testing.

Procedures

The teachers provided the main source of data. They were asked through the Maths Unit in the Ministry of Education to submit one copy of the test which they had administered to the students at the end of Term 3, 1995. They were also asked to provide the researcher with all the students' marked answer papers, all objectives or sources of objectives they used to construct the tests, and a copy of their test plans. The teachers completed a small questionnaire which provided additional data (see Appendix A).

Two officials from the Ministry of Education in the Seychelles were consulted. They provided additional background information about the teachers and the Ministry's policies regarding assessment in schools. The background information the

officials provided relates to the teachers' job descriptions and was applicable to all teachers at Primary 5A level.

Data Analysis

Responses to Questionnaire

The questionnaire sought background information about the teacher. All responses of teachers and that of the Ministry officials were summarised to obtain a background of the teachers' involvement in test construction.

Statistical Analyses of Tests

The following analyses were performed on the tests: calculation of the internal consistency reliability, equivalent form reliability, standard error of measurement, content validity assessment, and item analyses.

The statistical program EdStats (Knibb, 1995) was used for all statistical analyses performed on the tests. EdStats is a useful program that can be used to perform many analyses on test results and other statistical analyses in the social science area. The statistical package Minitab (1995) was used for drawing the histograms.

Results of all students (N = 20, except for Test F, I, and J where N = 16, and for Test G where N = 15) were included in the analysis. Stanley and Hopkins (1972, p. 269) pointed out that when item analysis or any statistical analysis is performed on test results, all the results should be included because economy of time is not an important consideration.

Internal consistency reliability. Cronbach's coefficient alpha was calculated for each of the test results. In cases when a test consisted of two papers (7 times out of

11), results of each paper were treated separately. The alpha values were then averaged to obtain a mean value for the set of tests submitted.

Equivalent form reliability. The four teachers who administered a two-paper test to their students combined the students' scores of both papers to obtain a composite score for each student. This practice of combining scores assumes that the two papers measure the same ability. Equivalent form reliability was calculated for all the two-paper tests to determine the extent to which the students' scores on one paper relate to their scores on the other paper.

Standard error of measurement. Standard error of measurement was calculated for each of the test's results. In cases when tests consisted of two papers, standard error of measurement was calculated for each set of scores. These values were then averaged to obtain an overall mean standard error for the tests submitted.

Item analysis. Difficulty and discrimination indices were calculated for items of all the papers. The results are reported as a percentage of effective, defective, and revision items. Histograms were drawn to illustrate the distribution of items over the range of difficulty and discrimination indices. The effective and the defective items were examined to ascertain the reasons for their effectiveness and their defects.

Content-Related Validity Evidence

Two judges rated the content-related validity evidence of the tests using a 10-point scale. A rating of 1 indicated very low content validity and a rating of 10 indicated very high content validity. The judges' ratings were correlated to determine the degree of inter-rater consistency then averaged to obtain a mean rating for each test.

The list below contains the criteria the judges used to assess a particular test.

1. Do the items match the objectives of the course?
2. Do all items belong to an particular objectives?
3. Are all objectives represented on the test?
4. Are the weightings of the items specified, and if so, are they fair?
5. Are there evidence of planning and preparation in the test construction?

CHAPTER 5

The Results

Overview of the Chapter

The first section gives an account of how the targeted teachers responded to the invitation to participate in the study. The second section summarises the responses to the questionnaire. The third section presents an analysis of the results obtained when the internal consistency and the equivalent form reliability of the test results were calculated. The fourth and the fifth sections report the results of the content-related validity evidence of the tests and that of the item analysis respectively. The sixth section examines and analyses the effective and the defective items of the tests.

An Overview of the Teachers' Participation in this Study

The Teachers' Reaction

As stated in the previous chapter, all Primary 5A mathematics teachers in the Seychelles were invited to participate in the study. Of the twenty-three teachers, fifteen (65%) responded to the invitation. Of the fifteen who responded, eleven (73%) submitted the following: a copy of their test, the students' marked answer papers, sources of objectives their tests were based upon, and responses to the questionnaire. The other four (27%) submitted only a copy of their test and sources of objectives their tests were measuring. Only two teachers (13%) actually submitted a set of objectives from which their tests were constructed.

The teachers who did not submit objectives or plans of their tests, explained that it would take them too much time to write all objectives they were assessing. Instead, they stated the titles of the books they used with their class during the year. All the teachers reported completing the books, Mathematics Primary Five 5.1 and

Mathematics Primary Five 5.2, (Maths Unit, 1989), given by the Maths Curriculum Unit for using at Primary 5A level.

The Obtained Materials

Altogether fifteen tests, one per each teacher, were obtained for the study. There was an average of twenty students' answer papers per test. The tests are labelled as Test A, Test B through Test O, indicating the teacher who submitted it. Out of the fifteen tests, eleven (73%) consisted of two papers. The other four (27%) consisted of one paper.

The supplementary materials, namely copies of the books the teachers used, and details of Primary 5A syllabus, were obtained from the Maths Unit in the Ministry of Education.

A Summary of the Responses to the Questionnaire

The questionnaire was administered specifically to acquire background information about Primary 5A maths teachers in the Seychelles schools. The following summary includes responses given by eleven teachers and two officials from the Ministry of Education in the Seychelles.

Institutions where the Teachers Trained

All teachers who participated in the study completed their teacher training in the Seychelles. Since measurement training has never been included in both the pre-service and the in-service teacher programs, it is concluded that these teachers are not formally trained to construct tests.

Number of Years of Experience the Teachers have in Test Construction

The number of years the teachers have been involved in constructing maths tests for their classes varies from two years to eighteen years (see Table 1).

Table 1

Number of Years of Experience the Teachers have in Test Construction

Teachers	No. of years
A	3
B	4
C	8
D	2
E	2
F	3
J	3
H	2
I	4
J	18
K	6
Median	3
Mean	5
Range	16
Standard deviation	4.7

Frequency of Constructing Maths Tests

On average, the teachers reported constructing one maths test every two weeks for all their classes. The officials reported that the Ministry of Education

requests that school teachers regularly test their students' achievement and keep records of their progress.

Decisions to be Made From Results of the Tests Submitted

The teachers reported that the results of the tests they submitted would be used primarily to evaluate how well the students have attained the objectives set for Primary 5A level and secondly, identify top performers and weak performers in the class.

The students' scores on these tests would be combined with their scores on tests from other subjects; the total score would be used in streaming them into different class groups. One teacher reported that results of the test she submitted would be used in assessing her teaching.

Use of Other Instruments in Measuring Students' Performances

All teachers reported using only paper and pencil exercises as the instrument they used to measure their students' performances. These are in the form of formative tests, assignments and homework.

However, the Ministry of Education in the Seychelles encourages the teachers to use their own discretion as to how they will go about acquiring information about their students' performances. Informal methods like observation may be used but the formal mode of assessment is via paper and pencil tests.

The Teachers' Involvement in Other Subjects

Teachers teaching maths at Primary 5A level normally teach Creole and science to the class or other primary classes. Whatever subject they teach, it is their responsibility to organise and design testing for all their classes.

The Reliability of the Test Results

Two forms of reliability, Cronbach's coefficient alpha and equivalent form reliability, and standard error of measurement were calculated for each set of test scores. The values obtained are presented in Table 2; their means are presented in Table 3.

Internal Consistency Reliability

The Cronbach's alpha coefficients for all the test scores were above 0.7. The highest was 0.95 and the lowest was 0.73. The mean coefficient alpha for Paper 1 results was 0.89 and that of Paper 2 results was 0.88. When all the alpha values were combined, the mean was 0.89. This indicates that in general the test results were highly reliable. High internal reliability indicates that items of the test consistently measure the same ability (Stanley & Hopkins, 1972).

Standard Error of Measurement

Since the test results were used to make decisions about the students' performances, an indication of the possible range in which their true score would fall is important. Thorndike, Cunningham, Thorndike, and Hagen (1991) noted that any interpretation of the test scores must be made with acute awareness of the standard error of measurement. They added that in interpreting the scores of an individual, it is the standard error of measurement that must be kept in mind.

Despite their high internal reliability some test results had a large standard error of measurement. This was mainly seen in results of Paper 2. Six out of the seven sets of Paper 2 results (86%) had a standard error of measurement greater than 3.5. A large standard error of measurement (SEM > 3.4) was also seen in three of the four (75%) single paper test results.

Table 2

Cronbach's Coefficient Alpha, Equivalent Form Reliability, and Standard Error of Measurement of the Test-Results

	Cronbach's coefficient alpha		Standard errors of measurement		Equivalent form reliability
	Paper 1	Paper 2	Paper 1	Paper 2	For to two-paper tests
Test A*	0.95	N/A	4.8	N/A	N/A
Test B	0.92	0.93	2.3	4.1	0.81
Test C	0.94	0.91	2.4	4.0	0.63
Test D	0.87	0.93	1.8	2.1	0.11
Test E	0.85	0.81	2.2	4.6	0.77
Test F*	0.94	N/A	3.6	N/A	N/A
Test G	0.85	0.91	2.5	3.8	0.73
Test H	0.85	0.73	2.3	5.3	0.21
Test I	0.94	0.91	2.4	4.8	0.91
Test J*	0.88	N/A	2.5	N/A	N/A
Test K*	0.84	N/A	3.5	N/A	N/A

Note. N/A stands for Not Applicable. It is used in instances when a test consists of a single paper, indicating that a value for Paper 2 could not be computed. The asterisks indicate the single-paper tests.

All Paper 1 tests were marked out of 40 and Paper 2 tests were marked out of 60.

The tests marked with an asterisks were all marked out of 100.

Table 3

Means of Cronbach's Coefficient Alpha, Equivalent Form Reliability, Standard Errors of Measurement of the Test Results

	Cronbach's coefficient alpha	Standard errors of measurement	Equivalent form reliability
Paper 1	0.89	2.3	N/A
Paper 2	0.88	4.1	N/A
Single paper	0.90	3.6	N/A
All the papers	0.89	3.3	N/A
Two-paper tests	N/A	N/A	0.60

The mean standard errors of measurement for Paper 1 and Paper 2 was 2.3 and 4.1 respectively. The mean standard error of measurement when the standard errors of all the papers were combined was 3.3.

Equivalent Form Reliability

Equivalent form reliability was calculated for all the two-paper tests. This provides an extent to which the students' scores on the two papers relate. That is, the degree of consistency of the students' ranking on both papers.

As shown in Table 2, of the seven two-paper tests, four (57%) have an equivalent form reliability between Paper 1 and Paper 2 above 0.7. Equivalent form reliability of Test C papers is 0.63. The other two tests, Test D and Test H, (29%) have a very low equivalent form reliability between their two papers. Such low values indicate that the two papers were not measuring the same ability throughout.

Assessing the Content-Related Validity of the Tests

The researcher and his colleague labelled as Judge 1 and Judge 2 respectively, assessed the content-related validity of the tests. During the first step of the assessment, they worked together. They examined the objectives given by the officials from the Ministry of Education and ascertained whether they were related to the books the teachers used during the year. The examination revealed that they were related. Once this was done, the two judges worked independently.

The objectives of Primary 5A mathematics course fall into six strands namely, Number, Basic Operations, Measurement, Fraction, Shape, and Statistics. The judges matched the items of each test to the strand(s) they believed the items were measuring. Then they rated the strands of the tests out of ten, indicating the extent to which they believed the strands were being represented. Finally, they gave each test an overall rating out of ten indicating its degree of content-related validity evidence.

The overall questions that guided the judges throughout their decision-making process were: (1) given the books the students were supposed to have completed, the objectives of the course, and the particular tests, what can be said about the content representation and item sampling of the tests?; and (2) can it be said that a particular test adequately measures the students' performances?

Tables 4 and 5 summarise the ratings the judges allocated to each test. The overall rating of a test is not an average of the ratings allocated to its strands; it expresses the judges' overall impression about the content representation of the tests.

Table 4

Rating Judge 1 Allocated to the Tests

Tests	Number	Basic Operation	Measurement	Fraction	Shapes	Statistics	Overall Rating
A	5	5	4	2	3	3	4
B	6	5	3	4	3	3	4
C	4	6	2	2	2	2	4
D	3	3	3	3	3	3	2
E	1	2	2	2	1	1	1
F	5	5	3	4	2	2	4
G	5	4	4	4	5	5	5
H	3	3	4	4	3	4	4
I	2	3	3	3	3	3	3
J	4	4	4	4	4	4	4
K	3	2	2	4	3	3	3
L	4	6	5	4	4	4	4
M	3	2	2	1	0	1	2
N	1	2	1	1	1	2	1
O	1	0	2	2	1	1	1

Table 5

Rating Judge 2 Allocated to the Tests

Tests	Number	Basic Operation	Measurement	Fraction	Shapes	Statistics	Overall Rating
A	3	3	2	1	2	2	3
B	4	5	1	2	1	2	2
C	4	6	1	3	2	2	3
D	1	1	2	1	1	1	1
E	2	2	5	1	1	1	2
F	6	6	2	3	1	0	4
G	3	2	2	2	3	3	4
H	3	3	3	4	4	4	4
I	2	2	2	3	3	3	3
J	3	4	4	3	4	4	4
K	3	3	3	6	1	1	3
L	2	5	4	3	3	3	3
M	2	1	1	0	0	1	2
N	1	1	1	2	0	1	2
O	1	1	2	3	0	0	1

Calculating Reliability Between the Judges' Ratings

Pearson r correlation between the judges' ratings was calculated to determine the degree of inter-rater reliability.

Correlation between the ratings of tests for each strand. This analysis ascertained the correlation between the judges' ratings of tests for each strand (see Table 6). Table 6 shows that 4 times out of 6 (67%) the correlation between the two judges' ratings was at least 0.7. This indicates moderately high correlation between the way they rated the tests. Their ratings of tests for the strand measurement were weakly correlated ($\underline{r}$ = 0.39). Agreement between the way they rated the strand fraction was moderate ($\underline{r}$ = 0.57). The judges' overall rating of the tests however, showed a strong correlation ($\underline{r}$ = 0.79).

Table 6

Correlation Between the Judges' Ratings of Tests for each Strand

Strands	Correlation
Numbers	0.72
Basic Operation	0.83
Measurement	0.39
Fraction	0.57
Shapes	0.76
Statistics	0.73
Overall Rating	0.79

Note. The correlation values are rounded off to 1 decimal place.

Correlation of the judges' rating of strands within individual tests. A correlation coefficient was also calculated between ratings of strands within tests. This value indicates whether there was consistency between the two judges' opinions about the way the strands of individual tests were being represented. The correlation values are tabulated in Table 7.

Table 7 illustrates the following: eight times out of fifteen (53%) the two judges' ratings correlated moderately ($\underline{r} > 0.6$). Four out of fifteen times (27%) their ratings correlated poorly ($0.3 < \underline{r} < 0.5$). Three out of fifteen times (20%) their rating did not correlate at all ($\underline{r} = 0$).

The correlation values tabulated in both Table 6 and 7 indicate that there was some consistency in the way the two judges rated the content-related validity of the tests except for the three occasions when their rating did not agree at all.

Table 7

Correlation of the Judges' Rating of Strands Within Individual Tests

Tests	Correlation coefficients	Tests	Correlation coefficients
A	0.95	I	0.45
B	0.87	J	0.00
C	0.94	K	0.41
D	0.00	L	0.93
E	0.47	M	0.89
F	0.98	N	0.00
G	1.00	O	0.64
H	0.33		

Averaging the Judges' Ratings

The values presented in Table 4 and Table 5 were averaged to obtained an estimated value of the degree of content-related validity evidence of the tests. The results are summarised in Table 8.

Table 8

The Mean of the Two Judges' Ratings

Tests	Number	Basic Operation	Measurement	Fraction	Shapes	Statistics	Overall Rating
A	4	4	3	1.5	2.5	2.5	3.5
B	5	5	2	3	2	2.5	3
C	4	6	1.5	2.5	2	2	3.5
D	2	2	2.5	2	2	2	1.5
E	1.5	2	3.5	1.5	1	1	1.5
F	5.5	5.5	2.5	3.5	1.5	1	4
G	4	3	3	3	4	4	4.5
H	3	3	3.5	4	3.5	4	4
I	2	2.5	2.5	3	3	3	3
J	3.5	4	4	3.5	4	4	4
K	3	2.5	2.5	5	2	2	3
L	3	5.5	4.5	3.5	3.5	3.5	3.5
M	2.5	1.5	1.5	0.5	0	1	2
N	1	1.5	1	1.5	0.5	1.5	1.5
O	1	0.5	2	2.5	0.5	0.5	1
Mean	3.0	3.2	2.6	2.7	2.1	2.3	2.9

The values in Table 8 illustrate that in general the judges rated the content-related validity evidence of the tests very low. On average they were giving an overall rating of 2.9 per test. The most valid test received an average overall rating of 4.5. Only one test (7%) received this rating. Three tests (20%) received an overall rating of 4, three (20%) obtained an overall rating of 3.5, another 20% obtained a rating of 3, one test (7%) received a rating of 2, 20% received a rating of 1.5 and the remaining 7% obtained a rating of 1. Owing to the fact that the rating was made out of ten, it can be concluded that the judges believed the content-related validity evidence of the tests were low.

Item Analysis

Difficulty and discrimination indices were calculated for items of Test A through Test K. All together the eighteen test papers consisted of 759 items, an average of 42 items per paper. Based on the calculated indices, the items were classified as effective, poor or defective items.

For item difficulty, the effective items were those items with difficulty indices falling in the range 0.2 and 0.8 exclusive. The poor items were those items with difficulty indices either 0.2 or 0.8. These were the items needing revision. The defective items were those items with difficulty indices either above 0.8 or below 0.2. These items were either too easy or too difficult respectively.

For item discrimination, items with an index above 0.3 were classified as effective items, items with an index between 0.2 and 0.29 were classified as poor items, and items with an index below 0.2 were classified defective items.

Table 9 summarises the results of item analysis performed on the tests. For both the item difficulty and item discrimination, the percentages of items falling in the

three categories are reported. The difficulty and the discrimination indices for all the test items are appended (see Appendix B).

Item Difficulty

The values in Table 9 indicate that on average a test contained 61% of effective items, 5% poor items and 34% defective items. The overall percentage of effective items was relatively low. However, 33% of the tests had a high percentage of effective items. All of them had a percentage of effective items greater than 70. On the other hand, 28% of the tests had a very low percentage of effective items. Test D Paper 1, for instance, had only 20% effective items.

A further examination of the item difficulty indices showed that the defective items were mainly easy items. Figure 2, illustrates this point. When all the indices were combined, of the 297 items classified as defective 89% had a difficulty index above 0.8 and only 11% had a difficulty index below 0.2.

From Table 9 it is seen that in general, for the two-paper tests, Paper 2 consisted of more effective items than Paper 1.

Discrimination Indices

On the basis of discrimination indices, Table 9 indicates that on average there were 67% effective items per test. Some tests (22%) contained more than 80% effective items. In fact, all the tests contained at least 50% effective items. However, 39% of the tests had at least 30% defective items.

There were on average 22% defective items and 10% weak or poor items per test. Figure 3 illustrates the distribution of items over the range of discrimination indices. The results also indicate that for the two-paper tests, in 71% of cases, Paper 2 had a higher percentage of effective items than Paper 1.

Table 9

Percentage of Effective, Poor and Defective Items as a Result of Item Analysis

	Item difficulty			Item discrimination		
Tests	Effective items	Poor items	Defective items	Effective items	Poor items	Defective items
A (63)	78	6	16	76	10	14
B_1 (41)	46	12	42	71	7	22
B_2 (33)	73	12	15	73	21	6
C_1 (41)	61	7	32	78	5	17
C_2 (32)	75	9	16	81	3	16
D_1 (40)	20	2	78	52	10	38
D_2 (34)	68	3	29	88	0	12
E_1 (40)	30	10	60	55	12	33
E_2 (27)	56	0	44	56	19	25
F (96)	46	0	54	60	10	30
G_1 (41)	66	0	34	56	10	34
G_2 (43)	74	0	26	70	11	19
H_1 (48)	69	6	25	54	13	33
H_2(26)	69	8	23	50	15	35
I_1 (33)	67	0	33	88	3	9
I_2 (38)	82	0	18	84	11	5
J (47	47	10	43	57	11	32
K (36)	78	0	22	64	14	22
Mean	61	5	34	67	11	22

Note: 1 and 2 stand for Paper 1 and Paper 2 respectively. The means are rounded off. The number in brackets under the column tests, refers to the number of items for that particular test.

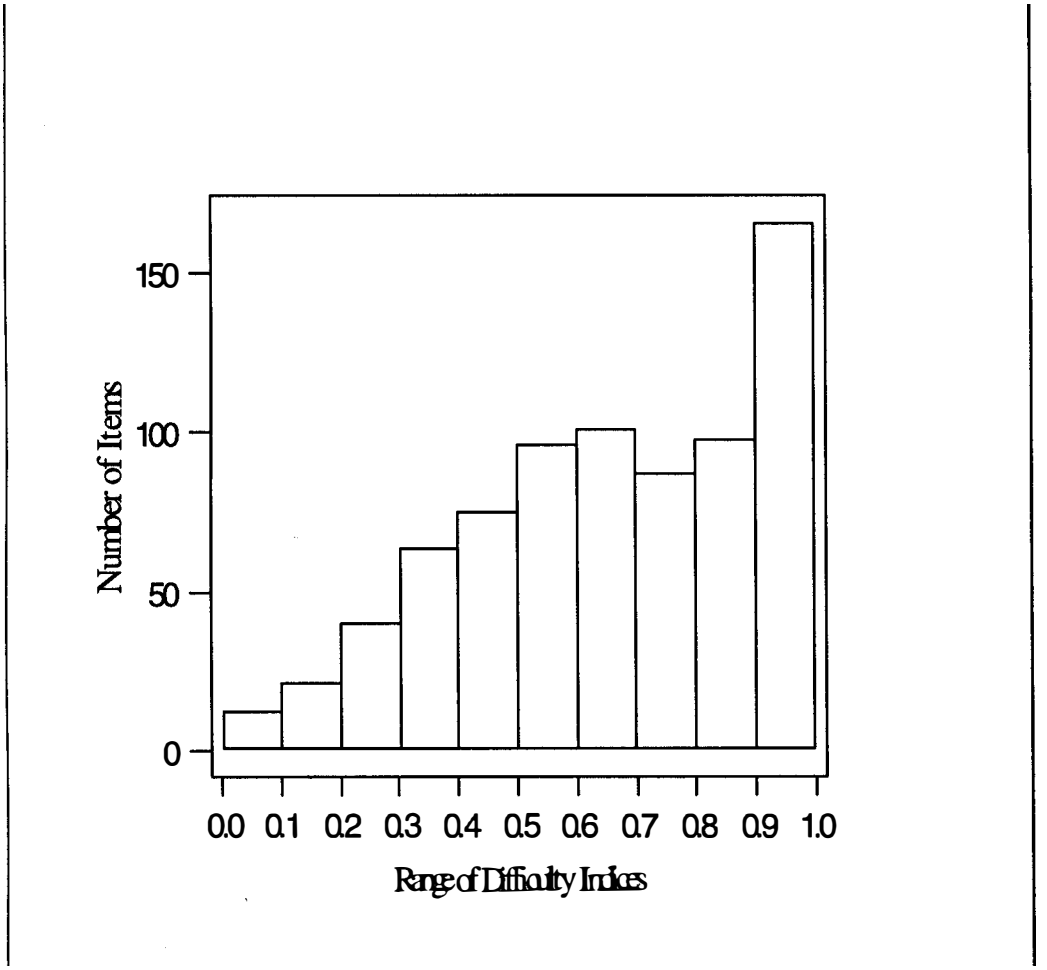

Figure 2. Histogram showing the distribution of items over the range of difficulty indices for all the tests combined.

Intersecting the Set of Difficulty and the Set of Discrimination Indices

The items were further examined to ascertain their effectiveness as indicated by both the difficulty and the discrimination indices. They were then classified effective or defective. The effective items were those items which were of appropriate difficulty level and at the same time positively discriminating between students. The defective items were those items which were rejected by both the difficulty analysis and the discrimination analysis (see Table 10).

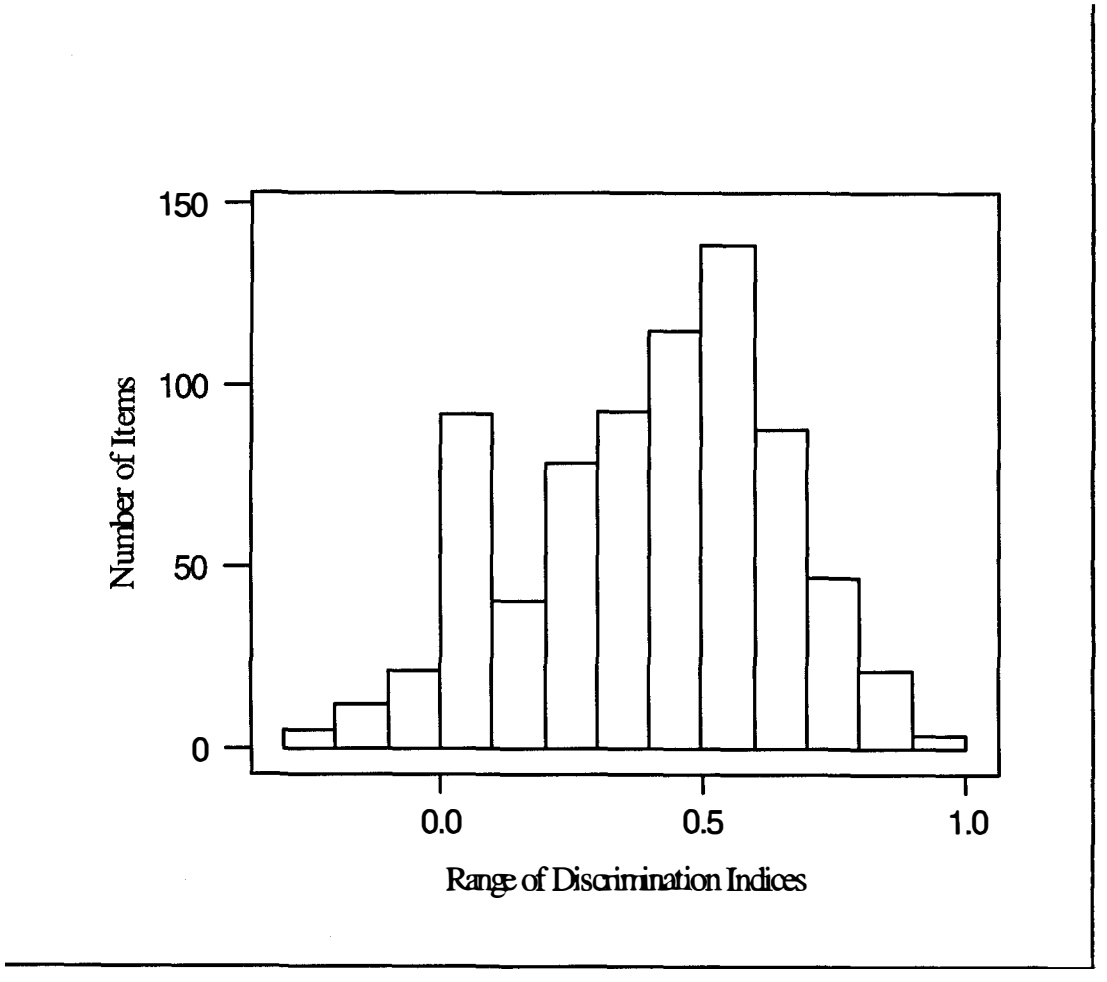

Figure 3. Histogram showing the distribution of items over the range of discrimination indices for all the tests combined.

Table 11 shows that on average a test has 50% very effective items and 16% defective ones. Eight out of eighteen test papers (44%) contained more than 60% effective items while nine other test papers (50%) had a percentage of effective items below the mean of all the tests. Test D, Paper 1, for instance, contained only 18% of effective items.

When the tests were classified by categories the following observations were made. On average the one-paper tests contained 50% of effective items and 19% defective items. For the two-paper tests, the percentage of effective items on Paper 2 was much higher than that of Paper 1.

Table 10

Percentage of Effective and Defective Items when the Indices were Intersected

	Effective items.		Defective items.	
Tests	Paper 1	Paper 2	Paper 1	Paper 2
A	68	N/A	10	N/A
B	46	61	20	6
C	61	65	15	6
D	18	62	38	6
E	23	41	30	22
F	35	N/A	25	N/A
G	46	60	20	16
H	42	38	15	12
I	61	68	6	3
J	40	N/A	23	N/A
K	58	N/A	17	N/A

Note. For a given test, the percentage values do not add up to 100. The difference relates to those items needing revision.

Table 11

Mean Percentage of Effective and Defective Items

Categories	Effective items	Defective items
One paper test	50	19
Paper 1 only	42	21
Paper 2 only	56	10
All papers combined	50	16

Characteristics of the Stand Out Items

The item analysis has provided valuable information about the teachers' competency in writing test items. Information has been gained about the type of items that they effectively constructed and the kind of items that caused concern. This section further examines these stand out items and reports their features. The stand out items refer to the most effective and defective items.

Features of the most effective items. In general these items varied from test to test and were not common to a specific strand. They were found mainly on Paper 2 of the tests or the second half of the single paper tests. Most of them were of the following form: (a) addition and subtraction of decimal numbers, (b) manipulation of mixed numbers and improper fractions, (c) calculation of volumes of cubes, (d) calculation of area and perimeter of shapes, (e) writing a given figure in words (and vice versa), and (f) interpreting data from a simple bar chart.

Features of the defective items. The items identified defective were common to most of the tests. In other words, they were the same type of items which were

causing problems on all tests. They were very easy items and were found mainly on Paper 1 and the first half of the single paper tests. Altogether 124 items (16%) were classified as defective. Out of these 124 items only 16 (13%) were difficult items.

Table 12 which shows how these items appeared on the tests, indicates the following:

1. Basic numeracy contained the highest percentage of defective items. Only two of the eleven tests (18%) did not have a defective item belonging to this strand.
2. The strands measurement, and fraction also contained a high percentage of defective items.

The most difficult items were fraction items and were of the following types: (a) calculation of a given fraction of a whole number, (b) shading a given fraction of a shape, and (c) addition of two fractions when their denominators are unequal. Some worded problems which require students to do more than one operation were also causing difficulty.

The strands shapes and statistics were not well represented on the tests and the teachers who included items belonging to these two strands in their tests, did so with little trouble.

Summary

This chapter has provided an insight into the quality of tests constructed by teachers without measurement training. Analysis of their tests shows that the tests produced high internal consistency reliability. For the seven two-paper tests analysed, the mean equivalent form reliability was greater than 0.7 except for two occasions when the equivalent form reliability was below 0.3.

Table 12
Number of Defective Items Belonging to each Strand of the Tests

	Strands of the tests						
Tests	Number	Basic Operation	Measurement	Fraction	Shapes	Statistics	Word Problem
A	1	2	1	4		1	1
B	1	4	4				3
C	1	4	2	3			2
D		7	1			1	
E	3		1	5	1		1
F	5	4	2	4			2
G	5	4	3	3	1		
H	4	5	5	4	1		2
I		1	1	2			
J	4	2	3	1			
K	2		2			2	1

The content validity of the tests was low. The two judges who rated the content validity of the tests gave on average an overall rating of 2.9 per test. Item analysis performed on the tests indicated that the teachers construct very easy items. These items were ineffective in discriminating weak achievers from top achievers on the tests. Consequently, the tests contained relatively low percentages of effective items.

Item analysis also shows that the problem items were common to most of the tests. Items belonging to basic numeracy were very easy items, while the difficult

items belonged to the strands measurement, fraction, numbers and one or two worded problems items.

Some tests consisted of two papers. Scores of Paper 1 tests were in general more accurate than those of Paper 2. However, items of Paper 2 tests were more effective.

CHAPTER 6

Discussion

Introduction

When tests are used to make decisions about students, the quality of the tests is of vital importance. When teachers are not trained to construct tests there is a reason to suspect and question the quality of the tests they construct.

This study aims at determining the quality of tests constructed by teachers without measurement training and ascertain whether their results are reliable and valid and can be used with confidence to make decisions about the students' learning. In order to provide answers to these questions, a sample of tests constructed by Primary 5A mathematics teachers in the Seychelles was directly analysed. In this study a measure of internal consistency and equivalent form reliability, content-related validity, and effectiveness of items were used as indication of the quality of tests. The present chapter discusses the results reported in the previous chapter.

The Reliability of the Test Scores

Internal Consistency Reliability

The internal consistency reliability for all the test scores was high ($r > 0.7$). This high internal consistency reliability may be due to the nature of the items. In general the items of the tests were objective. Consequently, the marking was objective. Objective items and objective marking assist in ensuring a high internal consistency reliability (Gay, 1991).

Other than the nature of the items and the marking techniques, possible explanations for the high internal consistency reliability in the test scores can be

explained by the length of the tests, and the characteristics of the students who took the tests.

The tests were relatively long. There was an average of 42 items per test paper. Ebel and Frisbie (1991) used the Spearman-Brown formula calculations to explain the theoretical relation between test reliability and test length. The effect of successive doubling of the length of an original five-item test, whose reliability was assumed to be 0.20 is shown in Table 13 below.

Table 13

Relation of Test Length to Test Reliability

No. of Items	Reliability
5	.2
10	.33
20	.50
40	.67
80	.80
160	.89
320	.94
640	.97
∞	1.0

Note. From Essential Of Educational Measurement (5th ed.), (p. 89), by R. L. Ebel and D. A. Frisbie, 1991, New Jersey: Prentice-Hall. Copyright 1991 by Prentice-Hall, Inc.

Long tests which have been carefully written tend to produce higher internal consistency reliability. This is because they provide an adequate sample of the behaviour being measured and the scores are apt to be less distorted by chance factors such as special familiarity with a given item or lack of understanding of what is expected from an item (Linn & Gronlund, 1995).

The group tested was heterogeneous. In the Seychellois schools placement in an A class is based on the students' combined scores of all subjects. Therefore, it is probable that the students' ability in mathematics varies significantly. Appendix C gives the standard deviation of the sets of scores. The standard deviations of the test scores were very large. The more heterogeneous the group is, the more reliable are its scores (Gay, 1991; Linn & Gronlund, 1995).

A further examination of the internal consistency reliability of the test scores revealed that the results of Paper 1 tests were more accurate than results of Paper 2. Results of Paper 1 tests had a mean coefficient alpha equal to 0.89 and a mean measurement error of 2.3 whereas results of Paper 2 had a mean coefficient alpha of 0.88 and a mean measurement error of 4.1. The difference in the accuracy between the results of Paper 1 results and that of Paper 2 can be explained by the nature of the items and the marking procedures. Items of Paper 1 comprised mainly simple basic arithmetic items. There was only one correct answer for a given exercise, thus the marking was done objectively. However, items of Paper 2 were somewhat subjective. The items required the students to do more computations and interpretations of worded problems. The teachers gave marks for the working. Even if students got the final answer correct, marks were deducted if working was not shown. There were many instances where marks were deducted for that reason. The large standard error of measurement explains that the teachers may have not been consistent in their

marking of Paper 2 items. This finding further supports the need for teachers to write objective rather than subjective items.

Equivalent Form Reliability

The equivalent form reliabilities between papers of the two-paper tests were also high. The mean reliability was 0.6. However, in two instances (27%) the reliability value was below 0.3. Ebel and Frisbie (1991) reported that if two papers measure the same ability the equivalent form reliability is expected to be around 0.7. The results show that in general the two papers did measure the same ability.

The Content-Related Validity Evidence of the Tests

The judges rated the content validity of the tests as very low. The average overall rating was 2.90 per test. The highest rating given was 4.5. Only one test (6.7%) received this rating. There were instances when the tests received an overall rating of 1. This illustrates that the judges observed little evidence of content-related validity in the tests.

Other features regarding the content validity of the tests that were also of concern were: the purposes of the tests were unclear. It was not easily recognised whether the tests were norm-referenced or criterion-referenced. The teachers did not write clear purposes that would indicate exactly what their tests were meant to be measuring.

Secondly, some strands were over represented. The strand basic numeracy, for instance, was over represented in all the tests. The tests contained too many simple basic operation items. As the difficulty indices indicate, these items were easy items. On the other hand, strands like shapes and statistics were under represented. Test F for example did not contain any item from this strand. Possible reason for this is that

items belonging to the strands shapes and statistics are more difficult to write so the teachers omitted them.

Thirdly, there were problems with the cognitive level of the items. The objectives of the mathematics course for Primary 5A expected that the students display knowledge, understanding, and application of what they have learnt. The items of the tests submitted for the study focused more on the lower level of the cognitive of domain of Bloom's taxomomy of educational objectives (Bloom, 1956). This is a consistent finding in teacher-made tests (Marso & Pigge, 1991; McMorris & Boothroyd, 1992; Oescher & Kirby, 1990).

Another weakness in the tests was the weighting of the items. The weighting of items relies on the teachers' judgment. There were no justifications why some items were weighted more than the others.

The low content-related validity evidence of the tests may be explained by the lack of planning in test construction. This point is supported by Marso and Pigge (1992) who concluded that a major reason why quality of teacher-made tests is poor is because of the lack of preparation in the test construction.

Tuckman (1988) described careful planning as one way of ensuring content-related validity evidence of a test. Gay (1991), and Linn and Gronlund (1995) agreed that the use of a table of specifications is essential to provide a base for careful sampling of test items. Without a definite plan, the teacher has no assurance that the relationship between objectives and items is established within the test. All the teachers who submitted the tests for the study did not submit a plan or a table of specifications for their tests construction. The teachers reported that it would take them too much time to write the objectives of the tests they submitted. This is further indication that the tests were not constructed with a plan. The imbalance of content was clearly noticeable in the tests.

The present study was not intended to find out why teachers do not plan their tests. It is believed, as noted by Marso and Pigge (1992) that lack of planning is due to lack of training. It may be probable that teachers are unaware of the ways they can plan their tests.

From the evidence gained through the background information about the teachers, it can also be speculated that the teachers have too many tests to prepare during the year. A mathematics teacher at Primary 5A level may also teach mathematics in other classes at this or other year levels and also teach other subjects usually science and creole, in other classes. Since they are responsible for organising testing for all their classes, it may be possible that the time available for preparing tests is insufficient.

Item Analysis

Although the literature provides acceptable ranges of discrimination and difficulty indices, it does not indicate the least percentage of effective items that would classify a test effective or ineffective. Nevertheless, it is clear that the higher the percentage of effective items the better the test. The tests analysed in this study contained at most 70% of effective items. In some instances the percentage of effective items was as low as 18% with at least 35% defective items. The results clearly demonstrate that the teachers were not always writing effective items. Effective items were those items that were of appropriate difficulty level and at the same time effectively discriminating weak achievers from strong achievers on the tests. When these effective items were further analysed to ascertain some of their features, it was found that they were the items that required students to do more thinking.

When a sample of the discarded items was re-examined to find out why they were ineffective, it was observed that these items were very easy. Items that are too easy are poor discriminators and thus defective (Thorndike, Cunningham, Thorndike, & Hagen, 1991). It was expected that end of year tests which are meant to identify weak and strong performers would not have a high percentage of very easy items. Easy items are normally included in the formative tests which measure the students' mastery on the topic they are learning (Gay, 1991).

The implication for test construction is that teachers should as far as possible avoid easy items on the tests. Instead, they should write items which elicit the students' higher order thinking. A justification of this statement can be seen by re-considering Table 10 of Chapter 5. Of the two-paper tests, 86% of the time Paper 2 contained a higher percentage of effective items. As argued previously this is due to the fact that Paper 2 contained more higher order items.

The study demonstrates that the teachers do not always write effective items. Item writing is a skill which is obtained through training (Roid & Haladyna 1982). Moreover, if the teachers are aware of the various methods used to analyse tests, they may improve the quality of the items they write.

Conclusions

This study provides valuable information about the quality of tests constructed by teachers without measurement training. In responding to the major research question; do tests constructed by teachers without training in measurement and testing produce valid and reliable results, the study shows that the tests may have high reliability, yet are low in content validity. As Linn and Gronlund (1995) pointed out, in interpreting and using reliability information, it is essential to remember that

reliability estimates refer to the results of measurement, thus a reliable test is not necessarily valid. Reliability is strictly a statistical concept (Linn & Gronlund, 1995).

In responding to the question; can results of these tests be used to make decisions about students' learning, the study indicates that teachers should be extremely cautious in using the results of these tests. The quality of the tests is far from ideal. For instance, the teachers wrote very easy items which failed to discriminate between students, and many of the objectives they taught in class were not evenly represented on the tests. That means the scores of their tests are not a fair representation of the students' ability. Therefore, the teachers are urged to use other assessment instruments to ensure triangulation to confirm faults and to gain valid baseline data for decision making.

In Chapter 3 of this thesis, the various factors which influence the quality of tests, namely knowledge of subject matter, knowledge of test construction, actual practice, time, resources, and teachers' work loads were discussed. This study paid particular attention to teachers' knowledge of test construction. The sample of teachers chosen did not have the required knowledge of test construction and this lack of knowledge was clearly noticeable in the sample of tests analysed.

The teachers did not plan their tests. This consequently lowered the content-related validity. As argued before, when tests are not planned or drawn using a table of specifications, there is no certainty that a fair balance of items can be established. Since the teachers have no knowledge of test construction, it is assumed that they cannot use the results of their tests to improve the quality of items constructed. Despite all these, it cannot be concluded that lack of knowledge of test construction was the only factor which lowered the quality of the tests analysed. The

other factors were not investigated. However, that the teachers have to construct many tests at one time and this may be a reason why they do not have enough time to prepare their tests.

Consequently, this study gives rise to further research questions which will be discussed latter in this section.

Limitations of the Study

There are a number of limitations inherent in this study. First, the sample of tests analysed was small. Secondly, the study considered only maths tests at Primary 5A level. Information about the quality of tests across other year levels and in other subjects areas is unknown. Moreover, the study has examined only one test constructed by each of these teachers. It cannot be claimed that the tests they constructed always resembles the picture received from this study.

Another limitation of this research is that the methodology used did not offer opportunity to fully ascertain the testing practices of Seychellois teachers.

Another weakness in this study is in regard to the procedures used to assess the content-related validity of the tests. The processes were somewhat subjective. If someone else was to judge the content-related validity evidence of the tests, it is possible that s/he would arrive at a different point of view. This decreases the reliability of the study.

Despite the small sample size of tests used in this study, the results clearly indicate that there exists a problem about test construction when teachers are not trained to construct tests.

Since the findings have been consistent over the teachers who participated in the study and the tests analysed, it is predicted that the findings can be generalised to a degree to the situation at other levels. Consequently, the findings of this study

may be generalised to other primary and secondary classes of the Seychelles and to other small developing countries where teachers have no or little training in measurement and testing.

Recommendations

Following the results of the study, recommendations will be made to the Ministry of Education, particularly to the School of Education in the Seychelles, to increase measurement literacy and eventually upgrade the quality of tests in the primary and the lower secondary schools.

The lack of content-related validity was speculated to be due to lack of preparation in test construction. Lack of preparation was predicted to be due to the teachers' ignorance of measurement knowledge. Poor quality items was argued to be due to teachers' lack of awareness about ways of improving test items. There is enough evidence to claim that there is a need to assist in-service teachers about test construction and increase their awareness of the fundamental principles of measurement and testing. This can be done by organising workshops for in-service teachers.

Teachers need guidelines on how to construct effective items. Some teachers were reluctant to submit their tests. This may be interpreted that they were aware that they are not professionally competent to construct tests.

The teachers should be encouraged to use alternative methods other than tests to measure their students' performances since there is evidence that the tests are not adequately measuring the students' performances.

There is a definite need to introduce a course in measurement and testing or its equivalent in the pre-service teacher training program.

Other educational institutions in the world who are not giving teachers measurement and testing training should also consider the results of this study.

Concluding Statement

The study has generated other research questions in the area of educational measurement and testing practices in Seychellois schools and small developing countries. These areas include: (a) the quality of tests at other year levels and in other subject areas; (b) the quality of tests versus the teachers' number of years of experience in test construction; (c) investigating whether there have been any negative consequences as a result of making decisions based on these tests; and (d) verifying whether the teachers are deriving the kind of objectives the Curriculum Unit would expect. There are also other areas of interest about testing practice in the schools that the school authority in the Seychelles may wish to investigate.

This study will be seen as a catalyst for future research into test analysis in the Seychelles and other countries with a similar educational situation.

References

Aiken, L. R. (1980). Content validity and reliability of single items or questionnaires. Educational and Psychological Measurement, 40 (4), 955-959.

Airasian, P. (1994). Classroom assessment (2nd ed.). New York: McGraw-Hill.

Bloom, B. S. (Ed.). (1956). Taxonomy of educational objectives: Handbook I: Cognitive domain. New York: David McKay Company.

Bloom, B. S., Madaus, J., & Hasting, T. (1981). Evaluation to improve learning. New York: McGraw-Hill.

Boothroyd, R. A., McMorris, R. F., & Pruzek, M. 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. (ERIC Document Reproduction Service No. ED 351 309)

Carey, L. M. (1994). Measuring and evaluating school learning (2nd ed.). Boston: Allyn & Bacon.

Carter, K. (1984). Do teachers understand the principles for writing tests? Journal of Teacher Education, 35, 385-397.

Crocker, L. M., Miller, M. D., & Franks, E. A. (1989). Quantitative methods for assessing the fit between test and curriculum. Applied Measurement in Education, 2 (2), 179-194.

Ebel, R. L., & Frisbie, D. A. (1991). Essentials of educational measurement (5th ed.). New Jersey: Prentice-Hall.

Farr, R., & Griffin, M. (1972). Measurement gaps in teacher education. Journal of Research and Development in Education, 7 (1), 19-28.

Fleming, M., & Chambers, B. (1983). Teacher-made tests: Windows on the classroom. In W. E Hathaway (Ed.), Testing in the Schools: New directions for testing and measurement (pp. 29-38). San Francisco: Jossey-Bass.

Frary, R. B., Lawrence, H. C., Weber, L. J. (1993). Testing and grading practices and opinions of secondary teachers of academic subjects: Implications for instruction in measurement. Journal of Educational Measurement: Issues and Practices, 12 (3), 23-30.

Frisbie, D. A. (1988). Reliability of scores from teacher-made tests. Journal of Educational Measurement: Issues and Practices, 7 (1), 25-35.

Gay, L. (1991). Educational evaluation and measurement (2nd ed.). New York: Macmillan Publishing Company.

Green, K. E. (1989). Measurement and research in the classroom: Directions for preservice education of teachers. In J. Braun, Jr. (Ed.), Performing teacher education: Issues and new direction (pp. 253-277). New York: Garland Publishing, Inc.

Green, K. E., & Stagers, S. F. (1986, April). Effects of training, grade level, and subject taught on the types of tests and test items used by teachers. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco. (ERIC Document Reproduction Service No. ED 269 430)

Griffin, P., & Nix, P. (1991). Educational assessment and reporting a new approach. Sydney : Harcourt Brace, Jovanovich Ltd.

Gronlund, N. E. (1981). Measurement and evaluation in teaching (4th ed.). New York: Macmillan.

Guilford, J. P. (1965). Fundamental statistics in psychology and in education (4th ed.). New York: McGraw-Hill Book Company.

Gullickson, A. (1984). Teacher perspectives of their instructional use of tests. Journal of Educational Research, 77 (4), 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., & Ellwein, M. C. (1985). Post hoc analysis of teacher-made tests: The goodness-of-fit between prescription and practice. Journal of Educational Measurement: Issues and Practices, 4 (1), 15-18.

Hopkins, K. D., Stanley, J. C., & Hopkins, B. R. (1990). Educational and psychological measurement and evaluation (7th ed.). New Jersey: Prentice-Hall.

Kim, J., & Mueller, C. W. (1978). Introduction to factor analysis: What it is and how to do it. Newbury Park: SAGE Publication.

Kirby, P. C., & Oescher, J. (1987, November). Testing for critical thinking: Improving test development and evaluation skills of classroom teachers. Paper presented at the annual meeting of the Mid-South Educational Research Association, Mobile, LA. (ERIC Document Reproduction Service No. ED 291 757)

Knibb, K. (1995). EdStats (Version 1.0.5K51) [Computer software]. Perth: Ken Knibb.

Kruskal, J. B., & Wish, M. (1978). Multidimensional scaling. Newbury Park: SAGE Publication.

Kubizyn, T., & Borich, G. (1993). Educational testing and measurement (4th ed.). New York: Harper Collins Publishers.

Linn, R. L. (1983). Testing and instruction: Links and distinctions. Journal of Educational Measurement, 20 (2), 179-189.

Linn, R. L. (1990). Essentials of student assessment: From accountability to instructional aid. Teacher College Record, 91 (3), 422-436.

Linn, R. L., & Gronlund, N. E. (1995). Measurement and assessing in teaching (7th ed.). New Jersey: Prentice-Hall Inc.

Marso, R. N., & Pigge, F. L. (1988, April). An analysis of teacher-made tests: Testing practices, cognitive demands, and item construction errors. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA. (ERIC Document Reproduction Service No. ED 298 174)

Marso, R. N., & Pigge, F. L. (1989). Elementary classroom teachers' testing needs and proficiencies: Multiple assessments and in-service training priorities. Educational Review, 13 (4), 1-17.

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

Marso, R. N., & Pigge, F. L. (1992, April). A summary of published research: Classroom teachers' knowledge and skills related to the development and use of teacher-made tests. Paper presented at the annual meeting of the American Educational Research Association, San Francisco. (ERIC Document Reproduction Service No. ED 346 148)

Maths Unit. (1989). Mathematics primary five 5.1 and 5.2. Victoria, Seychelles: Ministry of Education.

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. (ERIC Document Reproduction Service No. ED 350 348)

Minitab (version 10.5) [Computer Software]. (1995). State College, PA: Mintab Inc.

Newman, D. C., & Stallings, W. M. (1982, March). Teacher competency in classroom testing, measurement preparation, and classroom testing practices. Paper presented at the American Educational Research Association annual meeting, New York. (ERIC Document Reproduction Service No. ED 220 491)

Oescher, J., & Kirby, P. C. (1990, April). Assessing teacher-made tests in secondary math and science classrooms. Paper presented at the annual meeting of the National Council on Measurement Education, Boston. (ERIC Document Reproduction Service No. ED 322 169)

Oosterhof, A. (1994). Classroom application of educational measurement (2nd ed.). New York: Macmillan Publishing Company.

Roid, H. G., & Haladyna, M. T. (1982). A technology for test-item writing. New York: Academic Press Inc.

Satterly, D. (1981). Assessment in schools. Oxford: Basil Blackwell Publisher.

Sax, G. (1989). Principles of educational and psychological measurement and evaluation (3rd ed.). New York: Wadsworth Publishing Company.

Sireci, S. G., & Geisinger, K. F. (1992). Analysing test content using cluster analysis and multi dimensional scaling. Applied Psychological Measurement, 16 (1), 17-31.

Stager, S. F., & Green, K. E. (1984). Wyoming teachers' use of tests and attitudes toward classroom and standardized tests. Wyoming: Department of Educational Foundation and Instructional Technology. (ERIC Document Reproduction Service No. ED 252 575)

Stanley, J. C., & Hopkins, K. D. (1972). Educational and psychological measurement and evaluation. New Jersey: Prentice-Hall Inc.

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

Stiggins, R. J. (1994). Student-centred classroom assessment. New York: Macmillan College Publishing Company.

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

Thorndike, R., Cunningham, G., Thorndike, R. L., & Hagen, E. (1991). Measurement and evaluation in psychology and education (5th ed.). New York: Macmillan Publishing Company.

Tuckman, B. W. (1988). Testing for teachers (2nd ed.). Sydney: Harcourt Brace Jovanovich Publishers.

University of Cambridge Local Examination Syndicate. (1990). Assessing educational achievement in the Seychelles: A framework for the future. Report presented to the Ministry of Education, Victoria, Seychelles.

Wesman, A. G. (1971). Writing the test item. In R. Thorndike (Ed.), Educational measurement (pp. 81-129). Washington DC: American Council on Education.

Wise, S., Lukin, L., & Roos, L. (1991). Teachers' beliefs about training in testing and measurement. Journal of Teacher Education, 42 (1), 37-42.

Worthen, B., Borg, W., & White, K. (1993). Measurement and evaluation in schools. New York: Longman Publishing Group.

Appendix A

Questionnaire Administered to Acquire Background Information About the Teachers.

Dear P5A Maths-teachers,
Will you please complete this questionnaire as part of the study. Some items require that you provide a brief response while some require that you circle the most relevant alternative from the given list.
Your name is not required on this form.

1) Indicate the name of **all** institutions where you received your teacher training.

2) I have been involved in constructing mathematics tests for my students for approximately
   (A) less than one year. (B) between one year to five years.
   (C) between five years to ten years. (D) more than ten years.

3) How often do you construct maths tests for this class?

_____

_____

_____

4) What are the decisions that you will make from results of the tests you are submitting?

_____

_____

_____

Appendix A

Questionnaire Administered to Acquire Background Information About the Teachers.

5) Do you use other instruments for assessing the students' performances? (If yes, then list them.)

6) What other subjects (if any) do you teach ?

_____

_____

_____

7) Are you involved in preparing tests for all the subjects you listed in answer to question 6 above?

Yes No

Thank you for your time devoted in completing this form and your overall participation in this study.

Sincerely, Justin Valentin.

Appendix B

Results of Item Analysis

Table B1
Test A

Item no.	Item difficulty	Discrimination index	Item no.	Item difficulty	Discrimination index
1	1.00	0.000	28b	0.400	0.542
2	0.85	0.544	29	0.350	0.421
3	0.95	0.160	30	0.230	0.640
4	0.65	0.785	31a	0.350	0.649
5	0.85	0.136	31b	0.300	0.648
6	0.70	0.439	32a	0.625	0.502
7	0.65	0.603	32b	0.425	0.584
8	0.50	0.547	33a	0.250	0.523
9	0.60	0.370	33b	0.250	0.221
10	0.50	0.605	34a	0.700	0.439
11	0.70	0.436	34b	0.800	0.388
12	0.30	0.727	35	0.570	0.631
13	0.75	0.436	36a	0.400	0.578
14	0.45	0.690	36b	0.075	0.237
15	0.30	0.454	37	0.450	0.599
16	0.20	0.474	38	0.400	0.265
17	0.35	0.606	39a	0.950	-0.084
18	0.30	0.404	39b	0.150	0.607
19	0.30	0.185	39c	0.400	0.501
20	0.90	0.257	39d	0.950	-0.341
21a	0.70	0.717	39e	0.900	0.034
21b	0.65	0.760	40a	0.500	0.784
22	0.30	0.394	40b	0.600	0.676
23	0.50	0.793	40c	0.550	0.818
24	0.75	0.311	41a	0.450	0.705
25a	0.60	0.160	41b	0.400	0.688
25b	0.50	0.165	41c	0.250	0.282
26a	0.80	0.255	42a	0.350	0.576
26b	0.60	0.716	42b	0.300	0.764
27	0.80	0.494	42c	0.250	0.680
28a	0.70	0.659	42d	0.300	0.615
			42e	0.300	0.583

Appendix B

Results of Item Analysis

Table B2

Test B_1

Item no.	Item difficulty	Discrimination index	Item no.	Item difficulty	Discrimination index
1	1.000	0.000	22	0.600	0.175
2	0.800	0.255	23	0.900	-0.037
3	0.900	0.175	24	0.700	0.674
4	0.850	0.563	25	0.600	0.581
5	1.000	0.000	26	0.650	0.466
6	1.000	0.000	27	0.200	0.448
7	0.850	0.452	28	0.800	0.32
8	0.850	0.601	29	0.350	0.538
9	0.800	0.336	30	0.450	0.828
10	0.750	0.371	31	0.400	0.765
11	0.450	0.836	32	0.150	0.367
12	0.750	0.463	33a	0.350	0.688
13	0.950	0.221	33b	0.200	0.489
14	0.850	0.360	34	0.900	0.543
15	0.750	0.633	35	0.850	0.107
16	0.650	0.722	36	0.900	0.218
17	0.750	0.649	37	0.950	0.162
18	0.650	0.550	38	1.000	0.000
19	0.600	0.792	39	0.950	0.457
20	0.525	0.572	40	0.250	0.489
21	0.650	0.459			

Appendix B

Results of Item Analysis

Table B3

Test B_2

Item no.	Item difficulty	Discrimination index	Item no.	Item difficulty	Discrimination index
1	0.950	0.293	17	0.8	0.437
2	0.750	0.565	18a	0.750	0.582
3	0.600	0.672	18b	0.775	0.405
4a	0.675	0.437	19	0.675	0.283
4b	0.650	0.510	20	0.412	0.220
5	0.650	0.450	21	0.500	0.594
6	0.325	0.287	22	0.550	0.842
9	0.650	0.458	23	0.500	0.576
8	0.225	0.277	24	0.600	0.592
9	0.450	0.852	25	0.325	0.483
10	0.950	0.293	26	1.000	0.000
11	0.900	0.164	27	0.750	0.667
12	0.550	0.708	28	0.550	0.892
13	0.800	0.527	29	0.500	0.953
14	0.350	0.687	30a	0.850	0.384
15	0.450	0.482	30b	0.800	0.206
16	0.200	0.537			

Appendix B

Results of Item Analysis

Table B4

Test C_1

Item no.	Item difficulty	Discrimination index	Item no.	Item difficulty	Discrimination index
1	1.000	0.000	21	0.500	0.536
2	0.950	-0.191	22	1.000	0.000
3	0.850	0.134	23	0.900	0.063
4	0.900	0.407	24	0.850	0.302
5	0.600	0.511	25	0.700	0.587
6	0.550	0.446	26	0.800	0.443
7	0.750	0.556	27	0.650	0.396
8	0.850	0.472	28	0.850	0.104
9	0.500	0.742	29	0.500	0.719
10a	0.550	0.827	30	0.450	0.529
10b	0.450	0.878	31	0.550	0.675
11	0.900	0.426	32	0.650	0.597
12	0.800	0.457	33	0.550	0.560
13	0.600	0.305	34	0.600	0.733
14	0.950	0.277	35	0.450	0.714
15	0.950	0.277	36	0.550	0.594
16	0.850	0.519	37	0.450	0.854
17	0.750	0.651	38	0.450	0.854
18	0.800	0.113	39	0.550	0.722
19	0.350	0.430	40	0.550	0.526
20	0.700	0.488			

Appendix B

Results of Item Analysis

Table B5
Test C_2

Item no.	Item difficulty	Discrimination index	Item no.	Item difficulty	Discrimination index
1a	0.600	0.133	13a	0.625	0.424
1b	0.450	0.455	13b	0.200	0.234
2	0.550	0.593	14	0.275	0.581
3a	0.450	0.634	15	0.125	0.520
3b	0.250	0.709	16a	0.200	0.529
4	1.000	0.000	16b	0.350	0.378
5	0.425	0.675	17	0.950	0.126
6	0.400	0.691	18a	0.800	0.507
7	0.425	0.697	18b	0.550	0.733
8a	0.150	0.537	19	0.725	0.595
8b	0.150	0.537	20	0.775	0.474
9	0.300	0.355	21	0.500	0.683
10a	0.450	0.423	22	0.475	0.028
10b	0.450	0.650	23	0.425	0.086
11	0.325	0.499	24	0.475	0.475
12	0.300	0.692	25	0.325	0.741

Appendix B

Results of Item Analysis

Table B6

Test D_1

Item no.	Item difficulty	Discrimination index	Item no.	Item difficulty	Discrimination index
1	0.900	0.054	21	0.950	0.341
2	1.000	0.000	22	0.850	0.655
3	1.000	0.000	23	0.950	0.152
4	0.950	0.012	24	0.650	0.258
5	1.000	0.000	25	0.850	0.380
6	1.000	0.000	26	1.000	0.000
7	0.850	0.686	27	0.900	0.260
8	1.000	0.000	28	0.900	0.191
9	0.950	0.389	29	0.900	0.543
10	1.000	0.000	30	0.750	0.584
11	1.000	0.000	31	0.900	0.260
12	0.850	0.335	32	0.900	0.243
13	0.850	0.501	33	0.600	0.450
14	1.000	0.000	34	0.950	0.341
15	0.800	0.600	35	0.650	0.665
16	0.880	0.479	36	0.650	0.677
17	0.950	0.389	37	0.750	0.702
18	0.900	0.330	38	0.750	0.571
19	0.930	-0.109	39	0.900	0.507
20	0.650	0.593	40	1.000	0.000

Appendix B

Results of Item Analysis

Table B7
Test D_2

Item no.	Item difficulty	Discrimination index	Item no.	Item difficulty	Discrimination index
1	0.800	0.337	16b	0.750	0.655
2	0.350	0.301	17	0.500	0.532
3	0.900	0.322	18	0.750	0.584
4	0.450	0.831	19	0.850	0.462
5	0.700	0.473	20	0.575	0.819
6	0.625	0.478	21	0.550	0.870
7	0.625	0.675	22	0.525	0.846
8	0.850	0.574	23	0.500	0.677
9	0.425	0.114	24	0.850	0.509
10	0.950	0.181	25	0.600	0.329
11	0.950	0.556	26a	0.550	0.472
12	0.625	0.788	26b	0.350	0.685
13a	0.675	0.300	27i	1.000	0.000
13b	0.950	0.398	27ii	0.850	0.537
14	0.900	0.647	27iii	0.550	0.679
15	0.750	0.592	27iv	0.950	0.375
16a	0.550	0.568	27v	0.350	-0.026

Appendix B

Results of Item Analysis

Table B8

Test E_1

Item no.	Item difficulty	Discrimination index	Item no.	Item difficulty	Discrimination index
1	0.950	0.224	21	0.850	0.427
2	0.950	0.183	22	0.100	0.406
3	1.000	0.000	23	0.700	0.533
4	0.950	0.224	24	0.850	0.506
5	1.000	0.000	25	0.450	0.515
6	0.900	0.059	26	0.800	0.167
7	0.900	0.060	27	0.700	0.490
8	0.900	0.363	28	0.300	0.666
9	0.850	0.506	29	0.850	0.193
10	1.000	0.000	30	0.850	0.586
11	0.850	0.296	31	0.800	0.422
12	0.850	0.400	32	0.400	0.455
13	0.900	0.332	33	0.400	0.615
14	0.950	-0.225	34	0.850	0.427
15	0.950	-0.063	35	0.750	0.005
16	0.900	0.000	36	0.700	0.511
17	0.800	0.422	37	0.900	0.394
18	0.600	0.522	38	0.250	0.257
19	0.900	-0.030	39	0.450	0.515
20	0.800	0.236	40	0.550	0.500

Appendix B

Results of Item Analysis

Table B9

Test E_2

Item no.	Item difficulty	Discrimination index	Item no.	Item difficulty	Discrimination index
1/1	1.000	0.000	3iii2	0.100	0.099
1/2a	0.900	0.272	3iii3	0.925	0.032
1/2b	0.900	0.272	3iii4	0.383	0.501
1/3	0.300	0.576	4ia	0.417	0.520
1/4	1.000	0.000	4ib	0.417	0.500
1/5	0.400	0.266	5	0.850	0.344
2b	0.717	0.693	6a	0.650	0.548
2c	0.817	0.466	6b	0.650	0.348
3ia	0.900	0.333	7i	0.750	0.437
3ib	0.950	0.396	7ii	0.667	0.289
3iia	0.650	0.314	7iiia	0.700	0.404
3iib	0.450	0.050	7iv	0.400	0.395
3iii1	0.925	0.032	7v	0.750	0.203
			7vi	0.000	0.000

Appendix B

Results of Item Analysis

Table B10

Test F

Item no.	Item difficulty	Discrimination index	Item no.	Item difficulty	Discrimination index
1	1.000	0.000	34b	0.250	0.095
2	1.000	0.000	34c	0.312	0.438
3	0.812	0.354	34d	0.188	0.279
4	0.812	0.682	35	0.625	0.698
5	1.000	0.000	36	0.312	0.593
6	0.875	0.541	37a	0.875	0.757
7	1.000	0.000	37b	0.438	0.395
8	0.938	0.618	38a	0.938	0.618
9	0.938	0.212	38b	0.875	0.757
10	0.938	0.618	39a	1.000	0.000
11a	1.000	0.000	39b	0.812	0.073
11b	0.812	-0.022	40a	0.062	0.384
11c	1.000	0.000	40b	0.062	0.384
11d	1.000	0.000	41a	0.750	0.593
12a	0.875	0.137	41b	0.562	0.442
12b	0.875	0.328	42a	0.500	0.433
13a	0.875	0.328	42b	0.375	0.372
13b	1.000	0.000	43a	0.562	0.356
13c	0.938	-0.074	43b	0.688	0.421
13d	1.000	0.000	44a	0.625	0.494
14a	0.875	0.208	44b	0.812	0.781
14b	0.938	0.069	45a	0.688	0.581
15	0.625	0.410	45c	0.531	0.582
16	1.000	0.000	46	0.688	0.228
17	0.625	0.293	47	0.375	0.000
18	0.500	0.183	48	0.375	0.540
19	0.125	0.199	49	0.062	0.384
20a	1.000	0.000	50	0.750	0.554
20b	1.000	0.000	51	0.438	0.385

(table continues)

Appendix B

Results of Item Analysis

(continued)

Item no.	Item difficulty	Discrimination index	Item no.	Item difficulty	Discrimination index
21	0.625	0.504	52a	0.500	0.395
22a	1.000	0.000	52b	0.438	0.347
22b	0.938	0.405	53a	0.938	0.618
23	0.875	0.757	53b	0.938	0.618
24	0.750	0.264	54	0.562	-0.062
25	0.812	0.355	55a	1.000	0.000
26	0.750	0.270	55b	0.875	0.208
27	0.562	0.341	56	0.500	0.467
28	0.875	0.455	57	0.688	0.087
29	0.875	0.455	58a	0.688	0.617
30	0.562	0.031	58b	0.438	0.380
31	0.688	0.534	59b	0.562	0.601
32a	0.688	0.612	60	0.750	0.227
32b	0.938	-0.207	61	0.688	0.596
32c	0.938	0.212	62a	0.812	0.621
32d	1.000	0.000	62b	0.688	0.612
33a	0.750	0.461	63a	0.812	0.560
33b	0.812	0.318	63b	0.750	0.488
34a	0.562	0.423	64	0.500	0.324

Appendix B

Results of Item Analysis

Table B11
Test G_1

Item no.	Item difficulty	Discrimination index	Item no.	Item difficulty	Discrimination index
1	1.000	0.000	22	0.357	0.612
2	1.000	0.000	23	0.643	0.534
3	1.000	0.000	24	0.071	-0.125
4	0.786	0.800	25	0.000	0.000
5	1.000	0.000	26	0.500	-0.012
6	0.857	0.000	27	0.500	0.549
7	0.929	0.294	28	0.357	0.402
8	0.857	0.287	29	0.500	0.299
9	0.786	-0.314	30	0.714	0.414
10	0.714	0.319	31	0.429	0.450
11	0.929	0.340	32	0.929	0.575
12	0.929	0.575	33	0.643	0.469
13	0.929	0.340	34	0.000	0.000
14	0.714	0.595	35a	0.500	0.485
15	0.714	0.665	35b	0.500	0.226
16	0.714	0.497	36	0.714	-0.101
17	0.714	0.400	37	0.643	0.075
18	0.643	0.125	38	0.714	0.093
19	0.571	0.496	39	0.286	0.366
20	0.357	0.612	40	0.571	0.508
21	0.643	0.403			

Appendix B

Results of Item Analysis

Table B12

Test G_2

Item no.	Item difficulty	Discrimination index	Item no.	Item difficulty	Discrimination index
1	0.821	0.599	16	0.500	0.679
2a	0.500	0.439	17a	0.643	0.527
2b	0.643	0.603	17b	0.286	0.426
3a	0.429	0.383	18a	1.000	0.000
3b	0.286	0.306	18b	1.000	0.000
4a	0.000	0.000	18c	0.929	-0.248
4b	0.357	0.525	19	0.714	0.274
5a	0.643	0.280	20	0.071	0.073
5b	0.714	0.503	21a	0.679	0.628
6	0.357	0.459	21b	0.500	0.537
7a	0.857	0.351	22	0.286	0.408
7b	0.714	0.782	23a	0.571	0.414
7c	0.786	0.673	23b	0.571	0.292
8	0.786	0.206	23c	0.714	0.078
9	0.571	0.396	24	0.714	0.425
10	0.500	0.622	25	0.857	0.617
11a	0.714	0.782	26	0.286	0.591
11b	0.643	0.758	27	1.000	0.000
12	0.643	0.377	28	0.321	0.418
13	0.500	0.715	29a	0.143	-0.044
14	0.571	0.398	29b	0.071	0.208
15	0.286	0.662			

Appendix B

Results of Item Analysis

Table B13

Test H_1

Item no.	Item difficulty	Discrimination index	Item no.	Item difficulty	Discrimination index
1	0.950	-0.162	23	0.600	0.193
2	1.000	0.000	24	0.725	0.536
3	0.950	-0.162	25	0.250	0.456
4	0.950	0.234	26	0.575	-0.106
5	0.900	0.391	27	0.800	0.367
6	0.850	0.218	28	0.700	0.304
7	0.950	0.144	29a	0.400	-0.313
8	1.000	0.000	29b	0.425	-0.149
9	0.950	0.254	30a	0.900	0.151
10	0.750	-0.052	30b	0.800	0.497
11	0.800	0.356	31a	0.400	0.539
12	0.650	0.095	31b	0.400	0.641
13	0.550	0.799	32	0.875	0.342
14	0.700	0.413	33a	0.525	0.196
15	0.700	0.264	33b	0.500	0.272
16	0.600	0.445	34	0.750	0.547
17	0.650	0.593	35	0.650	0.282
18	0.750	0.537	36	0.250	0.308
19	0.950	0.355	37	0.700	0.225
20a	0.650	0.460	38a	0.300	0.448
20b	0.650	0.592	38b	0.300	0.448
21	0.750	0.580	39	0.650	0.436
22a	0.700	0.544	40a	0.500	-0.068
22b	0.400	0.512	40b	0.700	0.167

Appendix B

Results of Item Analysis

Table B14

Test H_2

Item no.	Item difficulty	Discrimination index	Item no.	Item difficulty	Discrimination index
1	0.300	0.031	12a	0.500	0.251
2a	0.900	0.381	12b	0.500	0.251
2b	0.800	0.542	12c	0.200	-0.006
3a	0.600	0.747	13	0.500	0.062
3b	0.400	0.529	14	0.300	0.523
4	0.425	0.025	15	0.600	-0.141
5	0.550	-0.019	16	0.650	0.676
6	0.900	0.120	17	0.400	0.476
7	0.450	0.367	18	0.925	0.351
8	0.650	0.682	19	0.100	-0.039
9	0.100	-0.030	20a	0.250	0.383
10	0.600	0.466	20b	0.850	0.279
11	0.450	0.282	20c	0.650	0.553

Appendix B

Results of Item Analysis

Table B15

Test I_1

Item no.	Item difficulty	Discrimination index	Item no.	Item difficulty	Discrimination index
1	0.938	0.309	16	0.188	0.506
2	0.750	0.672	17a	0.500	0.614
3	0.562	0.828	17b	0.250	0.505
4	0.688	0.701	17c	0.188	0.179
5	0.312	0.594	18a	0.438	0.823
6	0.750	0.624	18b	0.438	0.823
7a	1.000	0.000	18c	0.438	0.823
7b	0.750	0.217	19a	0.938	0.567
8	0.438	0.596	19b	0.938	0.567
9	0.625	0.585	19c	0.906	0.596
10a	0.500	0.558	19d	0.906	0.596
10b	0.312	0.699	20a	0.875	0.515
11	0.500	0.178	20b	0.938	0.567
12	0.375	0.390	20c	0.625	0.703
13	0.500	0.670	21a	0.688	0.656
14	0.250	0.648	21b	0.562	0.643
15	0.125	0.474			

Appendix B

Results of Item Analysis

Table B16

Test I_2

	Item	Discrimination	Item	Item	Discrimination
	difficulty	index	no.	difficulty	index
1	0.219	0.560	12a	0.812	0.539
2	0.156	0.537	12b	0.562	0.578
3	0.446	0.440	13	0.750	0.632
4	0.344	0.456	14	1.000	0.000
5a	0.688	0.706	15	0.438	0.590
5b	0.625	0.209	16a	0.438	0.850
6	0.719	0.689	16b	0.375	0.558
7	0.500	0.943	17a	0.781	0.431
8a	0.312	0.245	17b	0.825	0.409
8b	0.625	0.647	17c	0.438	0.850
8c	0.938	0.319	17di	0.562	0.731
9a	0.406	0.296	17dii	0.688	0.752
9b	0.344	0.530	18a	0.500	0.946
10a	0.375	0.276	18b	0.500	0.282
10b	0.250	0.163	19	0.281	0.685
11a	0.188	0.606	20a	0.688	0.788
11b	0.250	0.572	20b	0.188	0.490
11c	0.375	0.749	21a	0.562	0.697
11d	0.375	0.749	21b	0.500	0.775

Appendix B

Results of Item Analysis

Table B17

Test J

Item no.	Item difficulty	Discrimination index	Item no.	Item difficulty	Discrimination index
1	0.900	0.370	24	0.450	0.519
2	0.950	0.084	25	0.250	0.528
3	1.000	0.000	26	0.950	0.217
4	1.000	0.000	27	0.900	0.395
5	0.950	0.419	28	0.800	0.160
6	0.900	-0.019	29	0.600	0.569
7	1.000	0.000	30a	0.500	0.339
8	0.900	0.220	30b	0.300	0.635
9	0.800	0.591	31	0.700	0.334
10	0.350	0.728	32	0.300	0.686
11	0.800	0.289	33	0.650	0.466
12	0.750	0.335	34	0.050	0.151
13	0.650	0.580	35a	0.850	0.410
14	0.950	0.419	36	0.900	0.346
15	0.750	0.283	37a	0.600	0.427
16	0.950	0.117	37b	0.250	0.113
17	0.700	0.516	38	0.100	-0.019
18	0.900	0.222	39	0.150	-0.161
19	0.800	0.326	40	0.750	0.427
20	0.800	-0.181	41	0.700	0.669
21a	0.250	0.181	42	0.850	0.080
21b	0.250	0.422	43	0.650	0.745
22	0.650	0.307	44	0.400	0.411
23	0.900	0.198			

Appendix B

Results of Item Analysis

Table B17

Test J

Item no.	Item difficulty	Discrimination index	Item no.	Item difficulty	Discrimination index
1	0.900	0.370	24	0.450	0.519
2	0.950	0.084	25	0.250	0.528
3	1.000	0.000	26	0.950	0.217
4	1.000	0.000	27	0.900	0.395
5	0.950	0.419	28	0.800	0.160
6	0.900	-0.019	29	0.600	0.569
7	1.000	0.000	30a	0.500	0.339
8	0.900	0.220	30b	0.300	0.635
9	0.800	0.591	31	0.700	0.334
10	0.350	0.728	32	0.300	0.686
11	0.800	0.289	33	0.650	0.466
12	0.750	0.335	34	0.050	0.151
13	0.650	0.580	35a	0.850	0.410
14	0.950	0.419	36	0.900	0.346
15	0.750	0.283	37a	0.600	0.427
16	0.950	0.117	37b	0.250	0.113
17	0.700	0.516	38	0.100	-0.019
18	0.900	0.222	39	0.150	-0.161
19	0.800	0.326	40	0.750	0.427
20	0.800	-0.181	41	0.700	0.669
21a	0.250	0.181	42	0.850	0.080
21b	0.250	0.422	43	0.650	0.745
22	0.650	0.307	44	0.400	0.411
23	0.900	0.198			

Appendix B

Results of Item Analysis

Table B18
Test K

Item no.	Item difficulty	Discrimination index	Item no.	Item difficulty	Discrimination index
1	0.688	0.361	17ii	0.396	0.562
2	0.312	0.238	18i	0.375	0.726
3	0.812	0.003	18ii	0.438	0.541
4	0.812	-0.208	19/1	0.562	0.499
5	0.938	-0.245	19/2	0.562	0.499
6	0.500	0.511	19/3	0.688	0.708
7	0.875	0.319	19/4	0.562	0.530
8	0.469	0.631	20i	0.438	0.475
9	0.750	0.322	20ii	0.469	0.558
10	0.500	0.797	20iii	0.250	0.564
11	0.438	0.316	20iv	0.531	0.297
12	0.281	0.491	21	0.531	0.297
13	0.260	0.293	22b	0.917	0.046
14	0.328	0.358	22c	0.917	-0.037
15i	0.812	-0.161	23i	0.438	-0.013
15ii	0.750	0.484	23ii	0.438	-0.013
16	0.562	0.210	23iii	0.125	0.387
17i	0.646	0.440	23iv	0.312	0.416

Appendix C

Means and Standard Deviations of Test Scores

Table C1

		Means		Standard deviations	
Tests	No. of students who sat for the tests	Paper 1	Paper 2	Paper 1	Paper 2
A*	20	50.2	N/A	21.72	N/A
B	20	28.5	36.4	8.12	15.00
C	20	28.2	27.0	9.58	13.26
D	20	35.2	22.6	5.09	8.05
E	20	30.4	37.6	5.77	10.54
F*	16	69	N/A	14.01	N/A
G	15	24.4	34.0	6.38	12.64
H	20	29.3	31.2	5.96	10.16
I	16	25.3	31.9	9.90	15.97
J*	16	31.5	N/A	7.15	N/A
K*	20	31.0	N/A	8.55	N/A

Note. N/A stands for not applicable.

All Paper 1 tests were scored out of 40 and Paper 2 tests were scored out of 60. The tests marked with an asterisk were scored out of 100.