EXPLORING POTENTIAL OF RAISING VALIDITY EVIDENCE OF PHYSICAL SCIENCE TESTS THROUGH TEACHERS' PEER INSTRUCTION

PHD (TESTING, MEASUREMENT AND EVALUATION) THESIS

MIKE NKHOMA

UNIVERSITY OF MALAWI

CHANCELLOR COLLEGE

JANUARY 2012

EXPLORING POTENTIAL OF RAISING VALIDITY EVIDENCE OF PHYSICAL SCIENCE TESTS THROUGH TEACHERS' PEER INSTRUCTION

PhD (Testing, Measurement and Evaluation) Thesis

By

MIKE NKHOMA

M.Ed (Science Education) – Makerere University

Submitted to the Department of Foundation Studies, Faculty of Education, in fulfillment of the requirement for the degree of

Doctor of Philosophy (Testing, Measurement and Evaluation)

UNIVERSITY OF MALAWI

Chancellor College

January 2012

DECLARATION

I, the undersigned, hereby declare that this dissertation is my own original work which has not been submitted to any other institution for similar purposes. Where other people's work has been used acknowledgements have been made.

MIKE NKHOMA

_____

Signature

_____

Date

CERTIFICATE OF APPROVAL

The undersigned certify that this thesis represents the student's own work and effort and has been submitted with our approval.

Signature_____Date_____

Jimmy Namangale, PhD (Associate Professor)

Main Supervisor

Signature_____Date_____

Dixie Maluwa Banda, PhD (Associate Professor)

Member, Supervisory Committee

Signature_____Date_____

Dorothy Nampota, PhD (Associate Professor)

Member, Supervisory Committee

DEDICATION

To my family, relatives, friends and colleagues who inspired me in this task

ACKNOWLEDGEMENT

I am most grateful to my Supervisors, Dr. Jimmy Namangale, Dr. Dixie Maluawa-Banda and Dr. Dorothy Nampota for the role they played in this exercise up to April, 2010. The going was not easy but their guidance, encouragement and patience inspired me to come through it. I am also greatly indebted to Dr. Bob Chulu who, as Testing, Measurement and Evaluation expert, was always willing to provide resources and guidance. It is also fitting for me to gratefully recognize at this time encouragement and support received from Dr. and Mrs. Mvula.

I should also express my profound gratitude to The Malawi National Examinations Board for sponsoring and supporting this effort. In a special way, I should gratefully acknowledge the Executive Director, Mr. Mathews W. Matemba whose interest and support made it easier for me to accomplish the task. I also do acknowledge colleagues at the office for shouldering a lot of my office responsibilities, their words of encouragement and other services.

I am also most grateful to Mrs. Leah Kaira for her untiring effort to furnish me with reading material from the University of Massachusetts and Mrs. Jean Kayira from Washington and Saskatoon for the same. Lastly, I am most grateful to my family, relatives and friends for their encouragement and patience.

ABSTRACT

The study investigated the potential of raising validity evidence of Physical Science tests through teachers' peer instruction. Of particular relevance in the study was the effect of teachers peer instruction on content and construct related validity evidence of Physical Science classroom tests.

The study employed a mixed methods approach applying quantitative and qualitative methods. The quantitative methods used a pre-test and post-test one group experimental design, with peer instruction and validity evidence as independent and dependent variables respectively. Quantitative data was collected mainly from 62 tests with a total of 3016 items administered to 1543 learners of 2007 MSCE Physical Science class. The tests were constructed and administered by 17 Physical Science teachers to the class they were teaching. The teachers were purposefully sampled from Secondary Schools in the Southern Region of Malawi. Qualitative methods were applied to descriptive data generated through in-depth interviews with the teachers.

Item relevance, item representativeness and item cognitive representativeness in tests teachers constructed before they attended peer instruction were compared with item relevance, item representativeness and item cognitive representativeness in tests they constructed after peer instruction. Similarly proportion of shared variance due to common factors was the attribute of interest for comparison for construct related validity evidence in the same tests.

Content validity evidence of teachers' tests constructed after attending peer instruction increased significantly in terms of item representativeness. The proportion of shared variance in tests constructed after attending peer instruction also increased significantly, an indicator that construct related validity evidence had increased. The implication of the findings was that there is potential of raising the validity evidence of Physical Science tests through peer instruction.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF APPENDICES

ABREVIATIONS AND ACRONYMS

AERA	-	American Educational Research Association
APA	-	American Psychological Association
B Ed	-	Bachelor of Education
B Ed (Sc)	-	Bachelor of Education in Science
B Sc (Comp)	-	Bachelor of Science in Computer
B Sc (Eng)	-	Bachelor of Science in Engineering
BERA	-	British Educational Research Association
CFA_1	-	Common Factor Analysis
CFA_2	-	Confirmatory Factor Analysis
CPD	-	Continuing Professional Development
Dip Arch	-	Diploma in Architecture
Dip Ed	-	Diploma in Education
Dip Eng	-	Diploma in Engineering
EFA	-	Exploratory Factor Analysis
JCE	-	Junior Certificate of Education
KMO	-	Kaiser – Meyer – Olkin
M1	-	Mock Physical Science tests before 2007 (Pre-tests)

M2	-	2007 Mock Physical Science tests (Post-tests)
MIE	-	Malawi Institute of Education
MANEB	-	Malawi National Examinations Board
MERN	-	Manitoba Education Research Network
MSCE	-	Malawi School Certificate of Education
NAEP	-	National Assessment of Educational Progress
NCME	-	National Council on Measurement in Education
OSET	-	Office of Support for Effective Teaching
POST	-	Parliamentary Office of Science and Technology
PSLCE	-	Primary School Leaving Certificate of Education
SME	-	Subject Matter Expert
SPSS	-	Statistical Package for Social Scientists
SWED	-	South West Education Division
T1	-	End of Term 1 Physical Science tests (Pre-tests)
T2	-	Physical Science tests from the same domain as end of Term 1 Tests (Post-tests)

CHAPTER 1:

INTRODUCTION

1.0 Introduction

Chapter 1 is an introduction to the study. It has four main sections. The first section presents the contextual background to the problem. It is followed by sections on the statement of the problem, purpose and specific research questions and significance of the study. It ends with an overview of subsequent chapters of the dissertation.

1.1 Teachers' tests and formative assessment

Teachers' tests are a form of classroom assessment**,** whose purpose is to find out whether or not learners have benefited from instruction (Taylor & Nolen, 1996). Therefore the tests serve a formative role in instruction which does not only establish whether or not learners have achieved mastery of skills but it also guides instruction (Oosterhof, 2001). Teachers go over the tests in class with learners after giving them feedback to ensure that learners understand and master concepts and skills they could not demonstrate as a response to test items. Some teachers provide remedial lessons to individuals, groups of individuals, or even a whole class for that purpose. The subsequent teaching and learning

take into account observation made from the results of the tests. Therefore, classroom tests, as one form of assessment, are a tool for improving quality of teaching and learning leading to improvement of learner achievement, which is its ultimate goal (Boston, 2002; Black & William, 1998; Bude & Lewin, 1997). Thus the first role of teachers' tests is to serve a diagnostic function for improvement of instruction and learning leading to high learner achievement.

1.2 Teachers' tests and summative assessment

The second role of teachers' tests is that of reporting learning progress to the guardians or for school records. Assessment in this case is playing a summative role, which is a summary of learner achievement to date (Oosterhof, 2001; Thorndike, 1997). Therefore, teachers' tests in Malawi can serve dual roles of formative and summative.

Whatever the case, for the tests to be useful for their role in assessment, whether formative or summative, they must be properly designed, to be of good quality. They should be able to permit learners to demonstrate the breadth and depth of their knowledge (Shumway & Harden, 2003). Depending on their purpose, the tests must have the capacity to realistically reveal a broad range of learners' levels of thinking. Such information is useful to teachers for making accurate interpretations of learners' abilities and appropriate instructional decisions. All this amounts to saying that the tests must be of high validity.

1.3 Test validity

One attribute of a good test is that it has to be of high validity. Validity is the most important quality of a test (Gronlund, 1988). Definition of validity has been changing, over the time, with its conceptualisation. Shultz, Riggs and Kottke (1998) cite Garrett (1937, p. 324) who says that "the validity of a test is the fidelity with which it measures what it purports to measure"**.** This time around, different types of validity had been isolated, perhaps to enhance the clarity of the concept. The types of validity were predictive validity, content validity and construct validity. Predictive validity was most prominent than content and construct validity to the extent of Guilford (1946, p.429) saying "a test is valid for anything with which it correlates".

The turning point for the conceptualisation of validity was American Education Research Association (AERA), American Psychological Association (APA) and National Council on Measurement in Education (NCME) in 1985 endorsing that validity refers to the appropriateness, meaningfulness and usefulness of the specific inferences from test scores (Kang & Park, 2004). This definition, underscores concerns for validity to be with respect to evaluation of the quality of inferences based on test score (Guion, 1980). The endorsement of the new conception of validity marked a shift of attention for validity to 'evidence' when validity was defined as "the degree to which all the accumulated evidence support the intended interpretations of test scores for proposed purposes" (AERA, et al., 1999, p.11). Another critical aspect of validity as defined by AERA, et al. (1999) is establishing score meaning and use while Garret's definition is focused on how well a test is performing as a measuring instrument. The different foci about validity reflect conceptual differences about validity in the early days and now. Current consensus on

validity includes consequential basis of validity as proposed by Messick (1989a, p.13 in Linn, 1989), which leads to another possible definition of validity as "an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment". This definition projects a need for further refinement of validity as a concept. For purposes of this study, validity is defined as "the degree to which all the accumulated evidence supports the intended interpretations of test scores for proposed purposes" (AERA et al., 1999, p.11).

Validity is abstract. It is, therefore, determined in terms of logical and empirical information relating to a test as its evidence.

1.4 Quality of teachers' tests in Malawi

Concerns have been expressed about quality of classroom assessments in Malawi, which include tests. They are considered to be of poor quality, meaning that they are of low validity evidence. Mwanza and Kazima (2000) reported that recall and unfocused items dominate teachers' tests, and that the syllabus is not covered much in science in Malawi. Consequently validity in such tests is compromised. Johnson, Hayter and Broadfoot (2000) also reported similar findings of classroom assessments in Malawi.

Bregman and Bryner (2003) say that assessment in Africa, which includes Malawi, lacks cognitive depth. Similarly Kellaghan and Greaney (2003) reiterate this when saying that:

> There is evidence that the quality of teachers' assessment practices may be deficient in many ways. Problems that have been identified include the use of poorly focused questions, a predominance of questions that require short answers involving factual knowledge. (p.11)

The observations, therefore, are in line with the reports of Mwanza and Kazima (2000), and Johnson, et al. (2000).

1.5 Teachers' test construction skills

Poor quality of classroom assessment could be due to under-qualified or unqualified teachers teaching and assessing learners (Kellaghan & Greaney, 2003). Indeed employment of unqualified secondary school teachers is there in Malawi. Education statistics for 2006 show the magnitude of this problem (Ministry of Education and Vocational Training, 2006). Table 1.1 shows the statistics of 2006 secondary school teachers in Malawi for each qualification.

Table 1.1 2006 Secondary School teachers in Malawi

Qualification	Gender	No. of teachers	Percentage
PSLCE	F	11	0.34
	M	24	
JCE	F	6	0.56
	M	52	
MSCE	F	991	58.41
	M	5065	
Diploma in Education	F	588	20.68
	M	1556	
Diploma General	F	30	5.3
	M	520	
Degree in Education	F	219	9.75
	M	792	
Degree General	F	70	3.79
	M	323	
University Certificate of Education	F	12	1.17
	M	109	

Source: Education Statistics Ministry of Education and Vocational Training (2006)

In the context of this study suitable teaching qualifications for secondary level are Diploma in Education, Degree in Education and University Certificate of Education. Teachers with other qualifications would be unqualified or under qualified to teach in

secondary schools. In this study an unqualified teacher is defined as a person who is teaching a class of learners in secondary school but without having attained relevant teacher education qualification. In this regard, 68.4% of secondary school teachers in Malawi in 2006 as reflected in Table 1.1 were without appropriate teaching qualifications. As reported by Chavula (2008) in one of the local newspapers, the reaction of Ministry of Education to solve the problem through teacher recruitment in Malawi is as given in the text that:

> Those who intend to become secondary school teachers should brace for rigorous assessment as Ministry of Education has started administering interviews to applicants to reduce influx of unqualified teachers rocking the education sector, (…)
>
> Secondary schools have been beset with holders of Malawi School Certificate of Education and unemployed university graduates who have resorted to teaching due to lack of relevant employment. (…) It is disheartening to see people opening schools and becoming teachers without proper accreditation. (p. 3)

There are possibilities of teachers who are under-qualified to teach in secondary school. For purposes of this study an under-qualified teacher is a person who has attained some teacher education qualification but teaching a class of learners in secondary school in subjects which are not of their specialisation. This would include any person teaching a class of learners in secondary school but with primary school teaching qualification. Typical examples are a Bible specialist teacher teaching Physical Science or a primary school science teacher teaching Physical science in secondary school just to fill the human resource gap. This arrangement has the same effect on instruction as that of unqualified teachers. It increases the proportion of teachers who are not suitable to teach a class in secondary school from 68.4%.

Test construction skills for such teachers might be lacking, resulting in poor quality of tests and instruction in secondary schools. Test construction, in the context of this study, involves assembling of items in order to form a test. Table 1.2 shows the staffing situation for Agriculture, Biology, Mathematics and Physical Science teachers in South West Education Division (SWED), one of the six Education Divisions, in Malawi in the 2007 academic year. Some of the teachers are reported to be unqualified as defined in this study.

The large proportion of unqualified teachers who might be lacking test construction skills might translate into a substantial number of low quality teacher made tests being administered to those learning science in schools. The situation raises obvious concerns about the impact of low quality tests on instruction and learner achievement in science both at school and national levels.

Table 1.2 2007/2008 Secondary School Science teachers in SWED

Subject	No. Qualified	No. Unqualified	Percentage Qualified	Percentage Unqualified
Science and Mathematics	141	216	39.50	60.50
Physical Science	29	26	52.73	47.27

Courtesy of SWED by phone
Key: Science - Agriculture, Biology and Physical Science

The problem of teachers lacking skills for construction of classroom tests could be worse than envisaged. Other qualified teachers too might have forgotten what they learned in college about test construction or they might not have covered it at all. Kadzamira, Moleni, Kholowa, Nkhoma, Zoani, et al. (2004) found, according to members from some school communities they interviewed, that improper training of teachers was the possible cause of poor quality classroom assessment. Teachers are not properly trained in

assessment. They complete training with little or no skills for assessment. Therefore, the overall proportion of teachers lacking test construction skills could be more than what can be inferred from the statistics in Table 1.1 and Table 1.2.

Test construction skill deficiencies in schools seem to be real. In two independent studies secondary school teachers claimed to have little or no knowledge and skills for designing and carrying out assessments (Selemani-Mbewe, 2003; Kaira, 2003). It means that they could not write test items of good quality neither could they construct tests of high validity. The common practice in such a situation would be teachers copying past examination items for their tests (Chakwera, 2005), which does not guarantee validity of their tests. It is one thing to have items of good quality for a test. It is also another thing to assemble the same good items for a test of high validity. It calls for proper knowledge and skills of test construction to achieve that.

1.6 Problem statement

The research Department of the Malawi National Examinations Board (MANEB) observed that a majority of the teachers they involve in item writing had difficulties in producing good test items and assembling them for tests of good quality. Some of the teachers' test items were predominantly recall and sometimes the items were out of syllabus. This experience gave evidence to concerns about existence of low quality teachers' tests in schools. Therefore, the issue of concern in this study was the poor quality of teachers' tests for classroom assessment. The contributing factor for low quality of teachers' tests could be teachers' lack of test construction skills.

Therefore, to raise validity evidence of their tests, it is inevitable for such teachers to attend Continuing Professional Development (CPD) activities in test construction. A CPD is an on-going learning activity aimed at helping teachers to teach in more effective ways (MIE, 2008) and in the context of this study, in terms of test construction. The CPDs however, do use many instructional delivery strategies besides lectures. Therefore, the issue to be considered as well was a CPD strategy that could be used for effectively improving teachers' test construction skills in order to raise validity evidence of their tests in a cost-efficient manner. In this regard, the research question the study sought to answer was "What would be the potential of raising validity evidence of Physical Science tests through teachers' peer instruction?"

Therefore the aim of the study was to explore the potential of raising validity evidence of Physical Science tests by trying to improve, through teachers' peer instruction, test construction knowledge and skills of the teachers. Peer instruction in the context of this study, is learning in which learners 'exchange their personal views and test them against the ideas of others' as they build own knowledge (Southwest Educational Development Laboratory, 1995, p.2).

1.7 Purpose of the study: Research questions

The main purpose of this study was to explore the feasibility of raising validity evidence of Physical Science tests through teachers' peer instruction in test construction. By studying teachers' tests constructed prior and after their peer instruction, the specific research questions that guided the study were:

a. Were teachers' post-test items an equally relevant and representative sample of the test domain as pre-test items?

b. Did teachers' post-test items equally measure learners' cognitive ability levels as pre-test items?
c. Were the means of percentage total variances explained by common factors between the teachers' post-tests and pre-tests the same?
d. To what extent were teachers aware of the need for raising validity evidence of their tests?
e. What were teachers' perceptions about possibilities of raising validity evidence of their tests through peer instruction in test construction?

1.8 Significance of the study

The impact of low quality of classroom assessment is poor learner achievement. Therefore raising validity evidence of teachers' tests, might lead to improved instruction and high learner achievement.

An increase of validity evidence of teachers' tests through teachers' peer instruction, as observed from the results of the study, might mean that the same instructional strategy could be applicable for school based CPDs in test construction. The CPDs would be simple to organize and affordable for schools.

The study also contributes research-backed information in education on formative assessment in general, and teachers' tests in particular. It also provides guidance for future research.

1.9 Thesis outline

Chapter 1 is an introduction of the study. Literature review is covered in Chapter 2 while Chapter 3 presents the research design and methodology of the study. Results of the

findings for the study are presented and discussed in Chapter 4. Chapter 5 presents conclusions based on the findings, implications of the findings and recommendations for further research.

CHAPTER 2:

REVIEW OF RELATED LITERATURE AND RESEARCH

2.0 Introduction

Chapter 2 discusses validity theory which has evolved from a traditional to a unified conception. Covered in the discussion of validity theory are construct validation procedures, sources of validity evidence and statistical techniques for construct validation. The chapter presents reliability and item quality as well since they contribute to test quality. Peer instruction, which has been applied in this study, is also discussed in the chapter. Some studies on validity and peer instruction have been presented in the chapter to clarify some of the issues raised in literature review.

2.1 Validity theory

In testing, validity is a characteristic of a good test as discussed in section 1.2 on page 2 (Oosterhof, 2001; Thorndike, 1997; Guilford, 1946; Worcester, 1934). As stated earlier in Chapter 1 the concept of validity has been a subject of big debate. Efforts to clarify the concept, over the times, have resulted in distinguishing validity as a traditional conception on one hand and a unified conception on the other. A brief overview of validity theory is covered in this section.

2.1.1 Traditional conception of validity

Validity is a concept that stirred a lot of debate amongst early psychometricians, researchers and psychologists. Consequently, numerous notions of validity have been listed including face validity, predictive validity, concurrent validity, content validity and construct validity (Sireci, 1998; Cronbach & Meehl, 1955; Mosier, 1947). APA (1966) cited by Kang and Park (2004), after dropping face validity and other notions of validity as well as combining concurrent and predictive types of validity into criterion related validity, recognised the traditional category of three separate entities of validity which are criterion related validity, content validity and construct validity; a Trinitarian conception as Guion (1980) describes it. In the context of validity, a construct is "some postulated attribute of people, assumed to be reflected in test performance" (Cronbach & Meehl, 1955, p. 283). Criterion related validity was about how well a score on a test predicted a score on a criterion test, a test-criterion correlation (Guion, 1978). The assumption must have been that a person's attributes are constant, and therefore, a person's performance on a test and its criterion should be highly correlated. A good test, therefore, was supposed to show high correlation of a person's performance on a test and its criterion.

In another case, Tyler (1933, 1931) observed that a requirement for validity was the degree to which a test samples important objectives of the test domain. He was in a sense defining another type of validity referred to as content validity. Apparently, this was an expression of dissatisfaction with criterion related validity. Central to content validity was content representativeness of a test (Yallow & Popham, 1983; Guion, 1977). A good test, therefore, was expected to fairly sample the domain of interest.

Construct validity was assumed to apply to tests meant for measuring hidden psychological attributes of a person like intelligence, personality and anxiety for example

(Messick, 1989b; Shepard, 1993; Clark, 1959). The extent to which a test measured the psychological traits of interest determined the quality of the test. The dividing line between criterion related validity and construct validity was very thin. Both forms of validity were involved where measurement of a person's attributes was the issue but they differed in terms of procedures of how the attributes were measured under each of the forms. Under construct validity, a test was perceived to measure attributes of interest directly while in criterion-related validity, attributes of interest were predicted through correlation of a test and its criterion. In both cases, the issue was about the extent to which a test reflected to measure attributes of interest. Perhaps, this could be the reason why Garrett (1937, p. 324) cited by Shultz, Riggs and Kottke (1998) defined validity as "(…) the fidelity with which it measures what it purports to measure".

Construct validity was classified as an alternative procedure to criterion-related and content validity, besides ranking it lower than criterion-related validity (Angoff, 1988; Nunnaly, 1975; Cronbach & Meehl, 1955). Considering how validity is comprehensively conceived today there must indeed have been a serious debate about it over the period.

2.1.2 Weaknesses of criterion related and content validity

Many concerns were raised against criterion related validity and content validity in the traditional conception of validity. These forms of validity were not adequate for evaluating tests and that they were only simplified procedures for test validation (Sireci, 2007; Adcock & Collier, 2001; Anastasi, 1986).

Some of the specific shortcomings of criterion related validity include difficulties to obtain a consistent validity coefficient on repeated test validation and that external

influences to the scores, inherent in the testing process, are not taken into account (Clark, 1959). The other approach to this argument is that even a criterion itself, is amenable to items which are not relevant and representative of the domain of interest (Messick, 1989a in Linn, 1989). Kane (2001) elaborates this further by saying that:

> The criterion model does not provide a good basis for validating the criterion. Even if some second criterion can be identified as a basis for validating the initial criterion, we clearly face either infinite regress or circularity in comparing the test to criterion A, and criterion B, etc. (p. 320)

Therefore, validity of the original criterion might be difficult to ensure. Furthermore, when there is a correlation of a test and its criterion it could be for wrong reasons as well or they may not correlate at all if examinees are of the same abilities (Shepard, 1993). Hopkins (1998, pp. 97 - 99) illustrates the effect of lack of variability or range restriction on predictive validity coefficient. It is apparent that reduced variability leads to low validity coefficient.

Regarding content validity on its own is not sufficient for test evaluation. In the first place, it is not possible to achieve high content validity for a test because of mistakes made at construction (Guion, 1978). Supplementary information is necessary to appreciate validity of a test. Besides, content validation is used as a tool for establishing whether or not proper test construction procedures were followed (Kane, 1992; Guion, 1977). As another weakness of content validity, it may also not be possible to statistically achieve objectivity of test sample representativeness (Angoff, 1988). The concern in this respect is how sampling of important educational objectives should be done to achieve a high degree of sampling adequacy. Approached from the unified conception point of view, content validity does not directly support test score based inferences (Messick, 1989b). The

concerns about criterion related validity and content validity led psychometricians to seek a redress of the conceptual gap of validity.

2.1.3 Unified conception of validity

Lee Cronbach and Samuel Messick played an important role in developing the concept of validity using construct validity as its platform. Cronbach elaborated the evidential basis of construct validity while Messick extended it to include consequential basis of validity. Cronbach modeled validity on the positivist philosophy of science to bring it into the nomological network one of whose fundamental principles is to make clear what something is and in the context of tests, what a score means (Clark, 1959; Cronbach & Meehl, 1955). The shift of focus of validity to score meaning relative to the construct a test measures was intended to explain the behaviour a score summarises (Moss, 1995 & 1992). Validity which subscribes to score meaning is construct validity. Content and criterion related forms too, subscribe to score meaning. It follows that score meaning makes content and criterion related forms of validity be part of construct validity (Messick, 1995; Anastasi, 1986). The argument ushers in a unified conception of validity such that construct validity is the whole of the validity theory, where content and criterion related forms become its sources of evidence (Shepard, 1993).

Evidential basis of score interpretation and test use as expounded by Cronbach, reflected in AERA, et al., (1999) definition, did not address value implications of score meaning and social consequences of score use (Messick, 1989b). This is in consideration of the influence values have on score-based inferences and actions (Messick, 1989a in Linn, 1989). Messick, therefore, expanded the focus of construct validity to address both score

meaning and use with respect to value implications of testing and social consequences (Messick, 1995). This is well articulated in his 1989 definition of validity. As a result, unified conception of validity became a four-fold validity concept by taking into account both evidential and consequential basis of test interpretation and use (Shepard, 1997).

2.1.4 Construct validation

As a follow up to unified conception of validity all validation is also construct validation. It is a process for evaluating soundness of score based inferences (Guion, 1980 & 1978; Cronbach, 1971). Construct validation has adopted the same theory of verifiability of meaning, which is central to positivist's philosophy of science following efforts of Cronbach and Meehl (1955) in trying to establish what validity is. Construct validation is, therefore, an enquiry into score meaning. Initially, it used the positivist's method of enquiry but with time, several methods of enquiry have been recognized (Messick, 1989a in Linn, 1989). The foregoing girds construct validation to scientific enquiry, which leads to a systematic and methodological verification of score meaning. Perhaps it is for this reason that construct validation has a similar framework of a scientific enquiry, i.e. predict or hypothesise score meaning based on some theory, collect data to test the prediction or hypothesis and draw conclusions, on the results of the test, about the meaning of the score (Shepard, 1993). Kane (2002 &1992) describes this framework as an interpretive argument while Lane and Stone (2002) have called it a validity argument and the hypothesis or prediction, a proposition. Both are the same approach to construct validation using different terminologies.

Two issues come out clearly from this discussion. The first one is that construct validation framework outlines how inferences or interpretations derived from test scores should be validated. Therefore construct validation, by definition, becomes a procedure for validating score based inferences which confirms the argument that validity is not a property of a test (Sireci, 2007).

Secondly, construct validation assumes multiple sources of evidence as Cronbach and Meehl (1955) stated since score meaning can be arrived at from many sources of evidence. The sources of evidence are content, response process, internal structure also known as construct related source of validity evidence, relations to other variables and consequences of testing (Sireci, 2007; AERA, et al. 1999). 'Relation to other variables' is sub-categorized further into convergent and discriminant, test criterion relationship and validity generalization evidence (Kang & Park, 2004).

2.1.5 Sources of validity evidence

Sources of validity evidence are briefly described in this section. Content and construct related sources of validity evidence are discussed more than others because they were the basis for validity investigation in this study.

Content related validity evidence

Content related validation establishes evidence of whether or not test content is a representative and relevant sample of a defined content domain (Sireci, 1998a & 1998b; Yallow & Popham, 1983). Item representativeness assesses distribution of items in a test while item relevance assesses whether items come from within a defined test domain. Item

representativeness and item relevance are issues of construct under-representation and construct irrelevance respectively, which affect score meaning (AERA, et al., 1999).

The process of establishing content related validity evidence is judgmental. Subject Matter Experts (SMEs) form a panel of judges to review representativeness and relevance of test content (Crocker, Miller & Frank, 1989). As an example, in their study Valentin and Godfrey (1996) involved two SMEs as judges to establish content related validity evidence of teachers' tests in Seychelles. The judges rated the fit of the test in the content domain. However, the panel of two judges was not adequate. Besides, one of the two judges was the researcher himself, creating room for biased results. Therefore, the arrangement must have threatened the reliability of the results of content related validation. Sireci and Geisinger (1995) involved eight SMEs as judges for content validation in one study. In another case, Sireci (1998b) discusses a study he conducted with his colleagues on content validation of 1996 Grade 8 NAEP Science assessment. Ten science teachers were involved as carefully selected SMEs to rate the test. Another panel of ten SMEs was used by O'Neil, Sireci and Huff (2002) in their study of content validation of a State-mandated science assessment. Chakwera (2004) investigated validity of independently constructed curriculum-based tests using a panel of 15 SMEs. It is likely that the more the number of neutral SMEs involved in content validation the better for reliability of the results.

Another aspect of content related validation is rating scale applied in determining the fit between test items and content domain. Chakwera (2004) applied a 1-6 rating scale in the stated study. A rating of 1 represented 'not relevant and a rating of 6 represented 'highly relevant'. For cognitive skill relevance a 1 represented 'does not measure this skill'

and a 6 represented ‘it measures this skill very well’. Sireci (1998b) presents different approaches to rating for content validation. In the first approach, an item is assigned to one of the content areas or cognitive levels. There is no need for a scale in this context. However, his concern with this approach is that it provides information about content representativeness. Item relevance is not taken on board. A second approach applies a 10-point relevance rating scale. The mean rating across the SMEs gives the index of item relevance while the mean relevance rating across all items measuring the content area gives an index of content area representation. The longer the rating scale the more subjective it becomes. In this case, the difference in strength of opinion between some of the successive ratings would be too small to make a significant difference.

Sireci (1998b) and O’Neil, Sireci and Huff (2002) present another form of content validation using multidimensional scaling. The claimed advantage which this approach has over the afore-mentioned techniques, according to Sireci (1998b), is originality in rating since the content area is not given to SMEs. In this regard bias is reduced when assigning the items to a content area. Involvement of neutral SMEs as judges when rating scales are used would also reduce bias in content validation. The same could also be achieved when tests are made to be anonymous.

Response process

Downing (2003) presents a broad concept of response process as a source of validity evidence, which includes test administration, scoring, score processing and score reporting as its phases. The issue at stake about score meaning associated with response process is accuracy of data depending on the way each phase of the process has been managed. It can

lead to correct or wrong interpretation of test scores. Pedhazur and Schmelkin (1991) approach response process from the umbrella of logical analysis, which adds definition of construct, item content and method of measurement to response process. The additions show that logical analysis takes on board content related validity evidence for clarification of score meaning.

Messick (1989a in Linn, 1989) introduces the concept of substantive component of construct validity relating to response process. Its focus is the cognitive processes the respondent goes through which are a reflection of the targeted construct. It also adds to better understanding of score meaning. Cognitive processes relate to content related validity evidence in that items should sample all cognitive ability levels.

Internal structure

Internal structure source of evidence is also known as, construct related source of validity evidence. Central to internal structure of a test is the relationship of the test items and constructs they measure. A set of test items might measure the same construct, i.e. cluster around a construct, or not. Such information is useful for validity evidence. The relationship is reflected as item inter-correlation through internal structure analysis techniques (Shepard, 1993; Pedhazur & Schmelkin, 1991). Common Factor Analysis (CFA_1) is one of such techniques. It is a two stage analysis of Exploratory Factor Analysis (EFA) and Confirmatory Factor Analysis (CFA_2). EFA is applied to explore constructs thought to underlie responses to items while CFA_2 is a follow up to EFA, to verify its results (Froman, 2001). Therefore, the results of EFA are tested in CFA_2.

A number of issues require careful attention when conducting factor analysis. Sample size is one of them. Russell (2002) cites Comrey& Lee (1992) that the recommended sample size was a minimum of 5 or 10 participants per variable being analysed. Mundfrom, Shaw and Ke (2005) from their study had varied recommendations for minimum sample sizes in relation to variables proportionally including communality. When Russell reviewed some EFA studies published in Personality and Social Psychology Bulletin during 1996, 1998 and 2000, found that the recommendation for sample variable ration was not followed. There were fewer participants than recommended. MacCallum, Widaman, Preacher and Hong (2001) claimed that such rules were not valid. MacCallum, Widaman, Zhang and Hong (1999) in their study tool argue that the recommendation is not valid. They consider that the level of communality plays a critical role in factor recovery. They advise that communality must be high, at least 0.6. If it is low then the sample must be large. Concurring with them, Hogarty, Hines, Kromrey, Ferron and Mumford (2005) in their study too found that when communalities were high, sample size tended to have less influence on the quality of factor solution. Supporting them is Zhao (2008) who, based on his findings, concluded that the general rules of thumb of the minimum sample size are not valid and useful. He supports the idea that high communalities of at least 0.6 or mean value of communalities of 0.7 account for quality factor solutions. Field (2005) suggests communalities of at least 0.5 to be acceptable.

It is apparent in the foregoing discussion that when conducting factor analysis high communalities are critical for successful factor recovery. Attention should be given to sample size if the communalities are low. In such a case the sample size must be large. Sample size, as the only condition for EFA, would restrict construct validation studies to

classroom tests since it is not common at class level to attain recommended sample - variable ratio.

Besides sample size and communalities other issues to consider when conducting EFA are Kaiser-Meyer-Olkin (KMO), Bartlett's Tests of Sphericity and the determinant of the correlation matrix (Pedhazur & Schmelkin, 1991). These are measures of quality of data for factor analysis, meaning that they are indicators of whether or not EFA can be appropriately carried out (Field, 2005).

KMO measures sampling adequacy for factor analysis and it is expected to be greater than 0.5 in order to confidently carry out EFA on a given sample (Coughlin & Knight, 2007). It is used together with Bartlett's Test of Sphericity which should be statistically significant at $p < 0.05$ in testing appropriateness of correlations for factor analysis (Field, 2005).

A determinant is used to test for multicollinearity or singularity of a correlation matrix in factor analysis (Pedhazur & Schmelkin, 1991).The value of the determinant should be at least 0.00001 in order to properly extract the factors (Field, 2005) or else factor processing is terminated.

In carrying out factor analysis one must determine the model for factor extraction out of the several possible models. Costello and Osborne (2005) list unweighted least squares, generalized least squares, maximum likelihood, principal axis factoring, alpha factoring and image factoring as possible models for factor extraction. Principal axis factoring and maximum likelihood are the most common (Conway & Huffcutt, 2003; Darton, 1980).Principal axis is applicable in cases where distribution involving factor analysis does not assume normality while maximum likelihood does (Fabrigar, Wegener, MacCallum &

Strahan, 1999). Normality of a distribution becomes crucial usually when the intention is to test for significance of the outcome of EFA (Fabrigar, et al. 1999).

Regarding establishing number of factors to retain on extraction an option could be made from a number of rules again. They include Kaiser's extraction of eigenvalues greater than 1, Cattel's scree plot, minimum average partial correlation, Bartlett's chi-square test and parallel analysis (Tucker & LaFleur, 1991). Garson (2007) adds comprehensibility, variance explained criteria, Jollife criterion and mean eigenvalue. Kaiser's rule is the most commonly used (Henson & Roberts, 2006). Kaiser's extraction of eigenvalues greater than 1 is considered to be simple and objective (Fabrigar et al., 1999). Eigenvalue is a quantity which represents the amount of variance each factor accounts for (Taylor, 2004). Scree plot is also considered a useful method for determining the number of factors to extract (Field, 2005). This is a graphical presentation of factors against eigenvalues (Garson, 2007).

Consideration is also given to rotation of extracted factors in EFA. It is applicable in cases where more than one factor is extracted, to have a simple factor structure for easy interpretation (Russell, 2002). Therefore, rotation is concerned with improving factor solutions. Rotation procedures fall into two broad groups which are orthogonal and oblique (Newsom, 2007). Orthogonal procedures are applied when factors are assumed to be uncorrelated with each other while oblique rotation assumes that factors are correlated with each other (Darlington, 2007).

There are several of these rotation procedures under each category, orthogonal or oblique. Abdi (2003) identifies varimax, quartimax, and equimax as orthogonal procedures while Ender (1998) lists promax, maxplane, quartimax, oblimin and oblimax as oblique

rotation procedures. Fernandez (2003) and Costello and Osborne (2005) observe that orthogonal rotation procedures are more commonly used than oblique rotation procedures because they are simpler. In agreement, Abdi and Vicky (2009) say varimax is most commonly used of all the rotation procedures, meaning orthogonal procedure is popular. Fabrigar, et al. (1999), have a contrasting view. They recommend oblique rotation procedures on grounds that psychological attributes are expected to be correlated with each other and that oblique rotations provide more information than orthogonal rotations. Costello and Osborne concur with Fabrigar, et al., that although orthogonal rotation is commonly used, behaviour must be a function of correlated factors. Concurring with Fabrigar, et al., again, is Darlington (2007) who says oblique rotation procedures often achieve greater simple structure. Popularity of a rotation procedure does not necessarily mean that it is understood and correctly used.

Convergent and discriminant

Convergent and discriminant source of evidence is one of the external relationship sources of evidence. The key issue in this case is the score meaning on the basis of the relationship of test scores and scores from other tests (AERA, et al., 1999). The intention is to establish whether or not two tests measure similar or different constructs. Pedhazur and Schmelkin (1991) refer to the analysis of this source of evidence as cross structure analysis. The source of evidence is said to be convergent or discriminant if the relationship between test scores from two tests shows that the tests measure the same or different constructs respectively. Convergent and discriminant validation procedures apply multitrait-

multimethods matrix technique to determine the relationship between test scores from two tests (Downing, 2003).

Test-criterion relationship

Test criterion relationship is another example of external relationship source of evidence for a unitary validity concept. It applies the same procedures of test criterion related validation of the traditional validity conception. The test score remains the basis for predicting future performance (Shepard, 1993) which makes a test-criterion correlation coefficient to be the measure of interest. Therefore, the unified conception of validity takes on board the same concerns raised against criterion-related validity under traditional validity conception as discussed earlier in section 2.1.2. The degree with which the test predicts future performance on the criterion measure cannot be ensured (Pedhazur & Schmelkin, 1991).

Validity generalization

Validity generalization source of evidence uses results of past studies of test-criterion relationship validity evidence to predict performance in a new and same or similar setting (AERA, et al., 1999). For example, when predicting abilities in a future setting through aptitude tests, the assumption is that the tests would still be valid for testing similar or same abilities in the new setting. However, earlier arguments about weakness of criterion related validity evidence still hold. A test might be made to have multiple validity coefficients under different settings (Downing, 2003; Anastasi, 1986).

Consequences

In order to know the consequences of testing, scores must be put to use first. The information arising from use of test scores is the consequential validity evidence. Therefore, a consequential validation study involves investigating the appropriateness of intended testing purpose. It could also focus on unintended outcomes and effects, and adverse consequences which could be a result of test invalidity. As an example, Taleporos (1998) examining a testing situation in New York, is of the opinion that there were many intended and unintended consequences of testing in their public schools. Such consequences could be expected or unexpected. The expected might be planned for and therefore they are good effects. The unexpected may not be planned for and they could have either positive or negative effects. Lane and Stone (2002) cite a few cases of consequences of testing from several authors, which include improved learning for all students and narrowing down of the curriculum and instruction as intended and unintended effects respectively. Improved learning for all students would be planned for and therefore expected consequence of testing while narrowing down of the curriculum would be unplanned for and unexpected which would be really a regrettable consequence of testing. This would include teaching to examinations and rampant cheating in examinations experienced on the local scene as other examples of adverse effects of test score use.

2.2 Reliability

Reliability is briefly discussed in this study to clear any misconceptions which might be there since it is also a condition of quality of a test (Oosterhof, 2001; Thorndike, 1997; Guilford, 1947). Hopkins (1998) and Oosterhof take reliability to be an issue which

addresses the extent to which a test measures something consistently. What easily comes to light in this context is that reliability is concerned with possibilities of reproducing test scores on second administration of the test (Crocker & Algina, 1986). APA (1985) quoted by Pedhazur and Schmelkin (1991, p. 82) presents the classical aspects of reliability, "the degree to which test scores are free from errors of measurement". The foregoing is interpreted to mean that reliability should also be concerned with accuracy of the measure. Therefore test reliability addresses both accuracy and reproducibility of a measure.

The common ground for reliability and validity is the test score. While reliability is concerned with the accuracy and reproducibility of the score, the focus for validity is different. It is about the degree to which score based inferences are relevant for the purpose of testing (Thorndike, 1997).

Some of the situations which call for clarification of the relationship between reliability and validity, the question raised by Moss (1994), include whether or not a score would be relevant for the purpose of testing if it is precise but not consistent or if it is not precise but consistent or if it is both not precise and not consistent. Hopkins (1998, p.108) summarises the relationship between reliability and validity saying that "a measure can be reliable but may not be valid", and "a measure cannot have any validity if totally unreliable". Therefore, tests require moderate reliability for them to have some validity (Thorndike, 1997; Guilford, 1946). Recommendable moderate reliability ranges from 0.30 to 0.69 (Jackson, 2009). It is clear in this discussion that reliability is not a sufficient condition for validity but it is a pre-condition for validity (Oosterhof, 2001). Reliability, therefore, is another condition for validating a test besides conditions discussed in section 2.1.5.

Reliability is presented in several forms. Hopkins (1998) lists reliability coefficients computed for test-retest test, parallel or equivalent test, and single administration test forms.

Test-retest reliability coefficient, which is the coefficient of stability, is obtained through the computation of a correlation coefficient from scores obtained on repeated administration of a test to the same candidates at a later stage (Pedhazur & Schmelkin, 1991). Similarly a correlation coefficient can also be computed from equivalent forms of a test administered to the same candidates at two different times to obtain a reliability coefficient; coefficient of equivalence (Crocker & Algina, 1986). Equivalent form is also called parallel or alternate form. The time between one administration and the other is the major concern about test-retest and equivalent form procedures for estimating reliability. A new setting will have been created to influence test reliability.

Single test administration can be a solution to this problem. The reliability coefficient obtained in this case is a measure of internal consistency between the items, which indicates whether or not items of a test measure the same attributes (Thorndike, 1997). Split-half reliability coefficient is one of them, computed from two parallel forms of the same test comprising even numbered items as one form and odd numbered items as another (Hopkins, 1998).

Coefficient alpha (α) also known as Cronbach alpha is another single administration procedure for estimating reliability, computed using the equation given in Chapter 3 section 3.6.1. It is applicable for both dichotomously and polytomously scored items. Related to alpha coefficient is Kuder-Richardson which is presented as Kuder-Richardson 20 (KR20)

applicable only in cases where item scoring is dichotomous or Kuder-Richardson 21 (KR21) when the items are in addition assumed to be of equal difficulty (Hopkins, 1998).

2.3 Item quality

The quality of items contributes to the degree of test reliability and validity. An individual item can be assessed for quality, which is described in terms of its difficulty and ability to discriminate good examinees from the poor examinees.

The difficulty, also referred to as item difficulty, is expressed mathematically as an item difficulty index. For dichotomously scored items the difficulty index (p_i), is a proportion of examinees getting the item right, while for polytomously scored items it is a proportion of examinees' mean score on the item against its maximum score (Office of Education Assessment, 2009). Its computation applies the formula given in Chapter 3 section 3.6.1.

Item discrimination is the ability of the item to discriminate good examinees from poor ones (Office of Education Assessment, 2009). Literature has Pearson product moment, point biserial, biserial and phi correlation coefficients as some of the procedures for determining the degree to which an item discriminates good examinees from poor ones.

The Pearson product moment correlation coefficient, r, measures the degree of correlation between a score on an item and the total score on the rest of the items of a test excluding the score on the item (Crocker & Algina, 1986). It is computed using formula given in Chapter 3 section 3.6.1. It is a verification of whether or not an item measures the same attribute as the rest of the items in a test. It is applicable for tests whose items are either dichotomously or polytomously scored. Dichotomously scored tests can also employ

Point biserial correlation coefficient, which is a simplified form of Pearson product moment correlation (Kline, 2009). In cases where it is assumed that performance is an effect of an underlying attribute which is normally distributed a biserial correlation coefficient is applied (Crocker & Algina, 1986). Sometimes it is required to correlate performance on an item that is dichotomously scored with a criterion performance which is also dichotomous. In such a situation, phi correlation coefficient is applied (Young, 1999; Hinkle, Wiersma & Jurs, 1998).

2.4 Peer instruction

One other issue under consideration for literature review relates to CPD of teachers. It can be a tool for updating teachers' instructional skills so that learning is maximized (Hopkins, 1998; Training and Development Agency, 2008).

Central to the effectiveness of a CPD are appropriate instructional strategies for facilitating a participant's development of the targeted knowledge and skills. Traditionally, a lecture is popular for such activities. However, most of the lectures are intended for passing on information word for word to participants. The participants are told what to do, why and how to do it. Participants are therefore restricted to listening and note taking (Slavin, 2008). Such instructional strategies tend to be monologues with passive learners (Mazur Group, 2008). It follows that instructional strategies of this kind have the potential of promoting rote learning. It is perhaps in this context that instruction by telling is believed to be a flawed and less effective instructional method (Poulis, Massen, Robens& Gilbert, 1998; Thornton & Sokollof, 1998). McDermott (1993) highlighted several weaknesses of lectures which included their failure to:

1. integrate related concepts into a coherent framework

2. overcome misconceptions

3. develop learners reasoning abilities

4. provide a link among concepts, formal representations and the real world

Consequently meaningful learning could be less achieved because learners or participants are not intellectually engaged during traditional instruction. Peer instruction could be an alternative instructional strategy for enhancing teachers' functional understanding of concepts and procedures for constructing tests of high validity evidence.

2.4.1 Peer instruction: other definitions

Peer instruction is defined as an innovative technique that facilitates learners' engagement in the presentation material while enabling a dynamic, evaluative dialogue between the learners and the professor (Hillel, 2005). In this definition, however, for peer instruction to be useful it is the peer-peer interactions which should be stressed more than peer-instructor interaction. Learners would be at ease to express themselves to each other in demonstrating their reasoning and decision-making abilities. Peer instruction is also defined as a cooperative learning technique (Cortright, Collins & DiCarlo, 2005). Learners should work together, and also with their instructors in building knowledge.

Compared with traditional instruction, Kushnir (2006) defines peer instruction as an innovative tool that actively involves students in the instructional process. She uses the same qualifier, 'innovative', in her definition. Implied by 'innovative' is reform in instructional techniques for effective learning. The innovative instructional merit of peer

instruction perhaps is achieved as learners actively engage their mental faculties for further development (Meltzer & Manivannan, 2002). All these definitions reflect the potential which peer instruction might have for improving learning. As stated in Chapter 1 section 1.6, in the context of this study, peer instruction is learning in which learners 'exchange their personal views and test them against the ideas of others' as they build own knowledge (Southwest Educational Development Laboratory, 1995, p.2). The critical aspect of this instruction is each learner having some knowledge and skills for doing something; test construction as in the case of this study. This is very common in the teaching profession, where there is no expert instructor in a particular subject area.

2.4.2 Models of peer instruction

Crouch and Mazur (2001) describe a popularized form of peer instruction, Mazur's model. Learners read relevant course material before a class. The class is divided into a series of short presentations. After each presentation learners are given short questions to probe their understanding of the core concept. The learners individually answer the questions and present their answers to the instructor. Thereafter in small groups they discuss their answers. The instructor stops the discussion and explains the answer before moving on.

The issue of interest in Mazur's peer instruction is its shift from the traditional lecture and the impact it has on learning. The shift is the 'innovative' element sought in instructional techniques. Therefore, the question is whether or not a similar shift from a lecture would have the same impact in a CPD class for test construction.

The vital experiences of learners in peer instruction about how they arrive at the answers are stated by Heller, Keith and Anderson (1992) when they say:

> (…) students share their conceptual and procedural knowledge as they solve a problem together. During this joint construction of solution, individual group members can request explanations and justifications from one another. (p. 627)

It is such experiences that are critical for conceptual changes of learners, possibly for teachers as well in a CPD class for test construction. The experiences might be useful for developing functional understanding of concepts and procedures for constructing tests of much validity evidence.

Slavin (2008) tried peer instruction in an experimental demonstration. The procedure was the same as described by Crouch and Mazur except that experimental demonstrations replaced presentations. It might be an indicator of its potential to be of wide application.

Nicol and Boyle (2003) present a contrasting approach of peer instruction. Small peer group discussions of the concept question start, followed by an individual or group response, after which the students engage in a class wide discussion facilitated by the teacher. The order of events is reversed compared with Mazur's model of peer instruction. In addition, there is a class-wide discussion. While a class-wide discussion can be crucial in dissemination of important knowledge and skills from one group or individual to the other, it can also create room for other learners to contribute less in construction of knowledge on their own.

Lindboe (1998) has another form of peer instruction which involves low achieving learners of the same ability level. Some of them are given an opportunity to teach their peers materials that had been taught to them before and may not have been understood. In this

regard the peer teacher prepares, through investigation, the materials to be taught. This form of peer instruction concept is more of peer-tutor instruction.

Cuseo (2008) discusses the most common version of peer instruction where academically successful learners, advanced in their understanding of the subject matter provide learning assistance to the less advanced learners. In both cases of Lindboe and Cuseo elements of the traditional presentation method may not be eliminated. In fact they can be dominant. The peer tutor remains a source of knowledge. The interactive learner engagement, advocated for in peer instruction, would be curtailed.

The study models peer instruction described by Heller, et al. (1992) in which learners 'share their conceptual and procedural knowledge as they solve a problem together' about text construction. It is expected that 'during this joint construction of solution, individual group members can request explanations and justifications from one another' in the process fill knowledge and skills gaps they have about test construction.

2.4.3 Peer instruction: A constructivist instruction

Peer instruction is a highly interactive and learner centred instruction (Lasry, 2006). Learners interact with each other, the tutor and sometimes material for learning. Since learner centred instruction is also an attribute of constructivism, a peer instruction class is therefore a model of a constructivist class (Education Broadcasting Corporation, 2004). Constructivism is a philosophy of learning (Southwest Educational Development Laboratory, 1995). McDermott (1991) summarises how an individual would acquire knowledge in a constructivist class by saying that:

> An individual must construct their own concepts, and the knowledge they already have significantly affects what they can learn. The learner is not viewed as a passive recipient of knowledge but rather as an active participant in its creation. (p. 305)

Peer instruction and constructivism therefore have one thing in common. It is the emphasis on active learner engagement in building own knowledge based on their experience. Based on cognitive principles of learning, Redish (1994) like McDermott (1991) considers 'learners constructing their own knowledge' as the cornerstone of constructivism. Therefore, 'learners constructing their own knowledge' should also be a cornerstone of active learning in peer instruction, particularly in CPD classes for test construction.

Literature is abound acknowledging John Dewey, Jean Piaget, Levi Vigotsky and Jerome Bruner as people who contributed most to constructivism, as a philosophy of learning. They developed several psychological ideas and principles for promoting learning of children. Their ideas are similar in most cases. Dewey's philosophy was that education depends on action, knowledge grows from experience, and social interaction was important for the growth of this knowledge (Epstein & Ryan, 2002). Dewey's ideas are interpreted to mean that the learner has to be active in building own knowledge. Such knowledge is built on what a learner knows and a learner must interact with other learners and instructor to facilitate growth of the knowledge. Often times the ideal environment would be in small groups on their own to increase their social interaction, and making them more actively involved (Emand & Fraser, 2006). This is a thrust for learners to construct their own knowledge rather than simply receiving the knowledge from the instructor (OSET, 2008).

Piaget's cognitive constructivism focused on discovery as crucial for learning, because active involvement generates understanding (Epstein & Ryan, 2006). In discovery, the

learner has to interact with people and material to increase his knowledge and understanding. Interaction with people was the focus of Vygotsky's theories, for which he was called a social constructivist. He advocated social interaction and guidance as tools for learning (Riddle & Dabbagh, 1999). Bruner's ideas of active participation extend beyond the classroom. Learners need to participate in decisions about what, how, and when to learn (Epstein & Ryan, 2002). The significance of such ideas is that learners would own the instruction.

There are a number of issues to learn from the constructivists as a model of CPD classes. Interaction and experience is instrumental for learning. Therefore, an opportunity should be created for a learner to interact with learning material, other learners and the instructor. Interaction would keep a learner actively engaged in building own knowledge. The interaction should extend beyond the classroom as instruction is being planned to allow learners to have a voice on what they would like to learn. The other fundamental element of effective learning is the learner's experience in the subject matter because learning is building on previously acquired knowledge, which participants to a CPD should have in abundance.

Peer instruction as a mode of a CPD class for test construction would also draw important guidance from a recommendation of Southwest Educational Development Laboratory (1995) that:

> If the classroom can provide a neutral zone where students exchange their personal views and test them against the ideas of others, each student can continue to build understanding based on empirical evidence. Hands-on activities and observations of the natural world provide shared experiences for those constructions. (p. 2)

Peer instruction can apply such ideas to create a learning environment in which teachers in this study could freely explore each other's ideas for carrying out group tasks on test construction based on their experience and put them into practice. After practice they could assess each other's work to find out whether the ideas were correctly applied. This was envisaged to result in greater consolidation of development of test construction knowledge and skills even in the absence of an instructor.

2.4.4 Peer instruction: In adult education

Since participants for peer instruction in this study were adults the instruction can as well be considered to be adult instruction. Adult education is also learner-centred (Kerka, 2002). The claim makes adult education to be similar to a constructivist instruction and peer instruction in many respects. Therefore, principles of peer instruction and constructivism apply to it as well.

However, the way adults learn is slightly different from that of children (Williams, 2008). As a result, several other theories are applicable to adult education. One of them for example is that adults have to know why they need to learn something (Atherton, 2005). In the context of this study the adults were teachers. They might have been aware of their professional gap in test construction and the need to bridge the gap. Therefore, to them, attending the workshop must have been a search for solutions to the difficulties they had in test construction. A carefully planned peer instruction for them could have been more conducive for learning a lot from each other even without an instructor. It would also have created a good environment for self-directed learning. This is where Bruner's ideas of learner participation in planning instruction can be capitalized (Epstein & Ryan, 2002).

Cranton (1989) summarises postulates of Malcom Knowles, the architect of theories for adult learning, which in some ways might be different from the way children learn, as:

1. adults need to be involved in planning and evaluation of their instruction.
2. experience provides the basis for learning activities.
3. adults are most interested in learning of subjects that have immediate relevance to their job or personal life.
4. adult learning is problem centered rather than content oriented. (p.1)

Postulates 3 and 4 apply to adults more than to children. Peer instruction for a CPD class for test construction needs to integrate ideas of adult education and constructivism for it to be effective. The integration raises the potential more for peer instruction without an instructor.

2.4.5 Research in peer instruction

Much of the research on the effectiveness of peer instruction has been done in Physics classes. Hake (1998) compared the impact of interactive engagement versus traditional methods of teaching. Peer instruction is one of such interactive engagement methodologies. He surveyed 62 introductory physics courses that administered pre-tests and post-tests. Of the 62 courses, 14 used the traditional method of teaching while 48 used the interactive engagement method. A rough measure of the average effectiveness of a course in promoting conceptual understanding was based on the average normalized gain. The 14 courses applying traditional method had an average normalized gain of 0.23 + or – 0.04 while the 48 courses applying interactive engagement teaching method had an average normalized gain of 0.48 + or – 0.14. His conclusion was that interactive engagement method of teaching can increase mechanics-course effectiveness well beyond that obtained

in traditional practice. Interactive engagement in this context includes peer-peer interaction. The results are based on data from 62 institutions published or supplied through a questionnaire.

In a different study Crouch and Mazur (2000) applied a longitudinal design over a period of ten years to evaluate the effectiveness of peer instruction. Their source of data as a group was from two traditionally taught courses and eight taught through peer instruction in calculus and algebra based introductory physics courses. They administered pre-tests and post-tests at the beginning of instruction and end of term. Criteria of analysis were absolute and normalized average gain scores. Their results showed a greater increase of learner mastery of both conceptual reasoning and quantitative problem solving on the implementation of peer instruction than traditional method. Their result seems reliable in that it is based on more dependable data than that of Hake (1998).

A similar result was obtained by Fagen, Crouch and Mazur (2002) in their global survey to find out the success of peer instruction courses. Fagen, et al. (2002) solicited for data from instructors world-wide who used peer instruction in their courses. They assessed the gain of individual learners from colleges and universities that provided matched sets of pre-test and post-test Force Concept Inventory data. Most of the assessed peer instruction courses produced learning gains commensurate with interactive engagement pedagogies. It is not known how the data submitted for this study was selected. There is no guarantee, therefore, that the data received was not biased in favour of peer instruction.

Another significant result of peer instruction was obtained in Cegep by Lasry (2006) on replication of Mazur's model of peer instruction. Peer instruction enabled more conceptual learning than didactic lecturing. The design was experimental, involving a

control group and two experimental groups. The result shows that peer instruction can be applicable to other settings.

Cahyadi (2003) too, using an experimental and control group design in her study in Indonesia, to test effectiveness of interactive engagement as a teaching method in a physics course, found a similar result. The experimental group did better on conceptual understanding than the control group. However, in terms of learners' ability in problem solving she reports that it could not conclusively be shown by the examination score because of grading inconsistencies.

Cortright, et al. (2005) tested the hypothesis that peer instruction enhances meaningful learning. They defined meaningful learning as the learner's ability to solve novel problems or the ability to extend what has been learned in one context to new contexts. They divided their undergraduate class randomly into halves as control and experimental groups. During class, short presentations were given to both groups. Each presentation was followed by one-question multiple-choice quiz. Learners in the experimental group were allowed to discuss their answers with peers. The groups alternated after the first examination. Paired t-test results for significance in each case led to conclusions that experimental groups performed better during the examinations than control groups.

The results of the study of Cortright, et al. (2005) are significant in that the study was conducted in a physiology class. Peer instruction seems to be of effect even outside physics classes.

Based on studies discussed in this section peer instruction seems to have had a positive impact on an academic learning environment. The current study therefore, provided a different perspective under which the impact of peer instruction was being evaluated. The

question was whether or not peer instruction would have the same impact in a CPD environment for test construction. It was of interest to explore the aspects of test construction that could be amenable to improvement as a result of peer instruction.

2.5 CPD model in the education system in Malawi

Meke (2010) quotes Gray (2005) that CPD embraces the idea that individuals aim for continuous improvement in their professional skills and knowledge beyond the basic training initially required to carry out their jobs. The interpretation of this is that every professional undergoes basic training. Therefore, a CPD programme, in whatever form, is a requirement for any professional system. Realising this need, the education institutions in Malawi at all levels are engaged in CPD activities formally or informally for teachers' professional growth and development. To this effect, the Ministry of Education Science and Technology in Malawi has a CPD programme at Primary Level. One of its many aims is 'to enable teachers to develop skills, knowledge and understanding which will be practical, relevant and applicable to the classroom situation' (MIE, 2008, p. 2).

Perhaps the question is whether or not such aims of the CPD can be realised. Among other things, instrumental for realizing these aims in CPDs are the methodologies for instructional delivery. A brief discussion with the MIE office responsible for the CPD programme for the Primary School teachers revealed that the programme is a cascade model starting with national level to Division and District levels to PEAs at zonal level and down to the school level for generic needs. Schools too, identify local needs which are handled locally and can be shared with other schools in Clusters or Zones. Meke (2005) identifies cascade CPD models as transmissive; meaning an instructor of some sort has to

be involved to pass on knowledge and skills. The brief discussion showed that teacher to teacher sharing is also applied. It is also implied by MIE (2008, p. 1) in one of the meanings given for the CPD programme as 'ability to identify one's strengths and be able to offer expertise and share knowledge with others'. Therefore, the instructional strategy for the CPD at Primary Level combines both teachers' peer instruction and instructor based instruction. The issue of effectiveness of teachers' peer instruction applies in such CPD programmes.

2.6 Conclusion

The variables of this study were validity evidence as the outcome variable and peer instruction, as an independent variable. Validity is a sufficient condition of quality of a test. As a concept, it has evolved from a Trinitarian traditional conception to the current unified conception. In the current understanding of validity, what is validated is not the test but inferences based on test scores. Therefore validity is not a property of a test but score based inferences. In order to validate a test under unified conception of validity, sufficient evidence must be accumulated which is characteristic of a scientific enquiry, hence qualifying construct validation to be a scientific enquiry.

Peer instruction applies instructional strategies derived from constructivism and adult education. On the other hand adult education itself also applies theories of constructivism. The fundamental elements of peer instruction were building of knowledge and skills in test construction on their own, based on their experience and through learner active engagement with other learners and learning materials. Effectiveness of teachers' peer instruction in

CPD should also be an issue of interest to Ministry of Education Science and Technology. Peer instruction is applied in their CPD programmes too in Primary Schools in Malawi.

CHAPTER 3:

METHODOLOGY

3.0 Introduction

Chapter 3 describes the research methods, population and sample. Also discussed briefly in the chapter are school visits and peer instruction workshop in test construction. The chapter in addition covers data collection procedures and instrumentation together with related data analysis techniques. Research ethics which guided the moral conduct of the study have been presented as well. The chapter has also highlighted limitations to this study. It ends with a conclusion. The entire study has been summarized through a conceptual frame work.

3.1 Research methods and designs

The study was a mixed methods approach. It applied both quantitative and qualitative methods. The rationale was that quantitative and qualitative methods used together were ideal for collecting useful data for facilitating interpretation and in-depth understanding of whether or not improvement was made on validity evidence of teachers' tests for Physical science through peer instruction (Sydenstricker-Neto, 2005; Rocco, Bliss, Gallagher & Perez-Prado, 2003). Each method was useful for abstracting information which the other

could not. Therefore, the methods complimented each other, though with quantitative methods dominating.

Quantitative methods were applied in a pre-test post-test one-group experimental design in which peer instruction was an independent variable and validity evidence of teachers' tests a dependent variable. Pre-test post-test one group experimental design was preferable to pre-test post-test control group experimental design because the pre-test post-test control group experimental design was going to have more interfering variables to control. As a result it was going to be difficult in this study to equate groups on interfering variables. The rationale for quantitative methods was to collect numeric data in order to confirm hypotheses about the potential of raising validity evidence of teachers' tests in Physical Science (Neill, 2007).

In qualitative methods, a phenomenological design was used. The rationale for qualitative methods, and phenomenological design in particular, was to capture information from teachers' descriptions of what they experienced relating to test construction during this exercise in schools (Creswell, 1998). The information was useful for understanding issues involving improvement of quality of teachers' Physical Science tests from their perspective. Such contextual information could not be realized through quantitative methods. Therefore, the significance of qualitative methods in this study was to add, through probing, breadth and depth to the inquiry (Rocco, et al., 2003).

The conceptual framework of this study is summarised in Figure 3.1

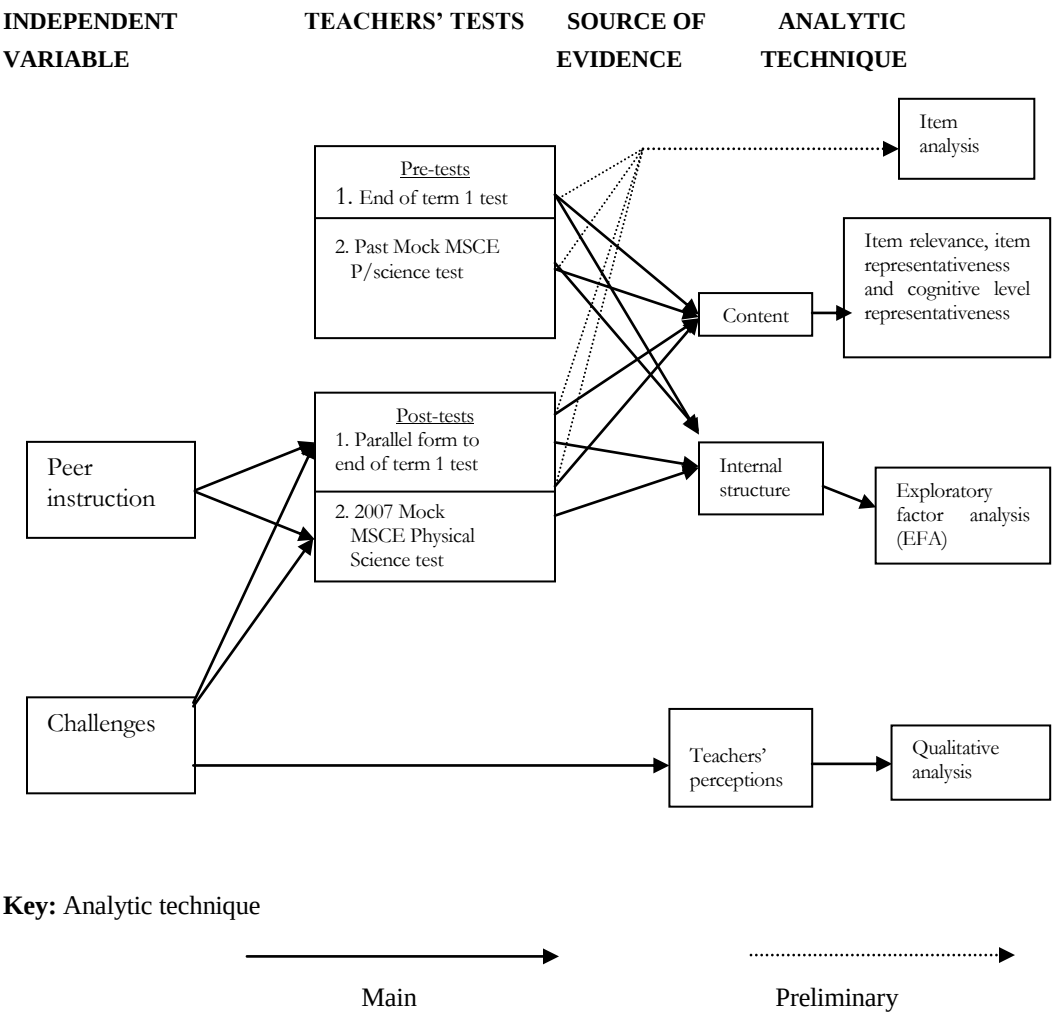


Figure 3.1 Conceptual framework of the study

The framework shows peer instruction which was the independent variable and validity evidence from the cited evidence sources as a dependent variable. Possible challenges were taken into account. They were considered as interfering variables in the study. They were to be identified through teachers' perceptions as constraints teachers experienced with test construction after peer instruction. The framework also provides a list of teachers' tests under consideration as well as sources of evidence, and analytical techniques for construct validation as applied in the study.

3.2 Population and sample

The sampling frame for the study was, teachers teaching 2006 Form 3 Physical Science class in conventional and private secondary schools in the Southern Region of Malawi. They must have taught a Form 4 Physical Science class and set an MSCE mock test previously. The assumption was that the 2006 MSCE Physical Science teacher was going to teach Physical Science to the same class in 2007 when it was going to be a Form 4 class. The teachers to be selected as a sample must not have been involved in test development activities with MANEB.

The sample plan was for 25 subjects. Targeting conventional and private secondary schools in the southern Region of Malawi was targeting 145 secondary schools. Basing on teacher characteristics described in the preceding paragraph, only 18 teachers qualified. As a result, they were all included in the sample within the demographic limits shown in Table 3.1, and qualifications and experience shown in Table 3.2. See Appendix 3.1, for sample profile. Sampling of teachers was therefore purposeful. Random sampling would have resulted in selecting teachers that did not have the characteristics of interest.

Table 3.1 Distribution of teachers from selected schools

	School type		
School classification	Boys only	Girls only	Co-education
Government Full Boarding	*	*	8
Double shift	*	*	3
Government Day	*	*	1
Grant aided (Boarding)	*	1	1
Private mission (Boarding)	*	*	2
Private circular (Day)	*	*	2
Total number of teachers	*	1	17

Key: (*) – no teacher was selected from this type of school

Table 3.2 Teachers' qualifications and experience

Qualification	Frequency	Experience (Years)	Frequency
MSCE	1	1	1
Dip Ed	8	2	5
Dip Arch	1	4	3
Dip Eng	1	5	1
B Ed (Sc)	4	7	1
B SC (Eng)	1	11	1
B Sc	1	12	2
B Sc (Comp)	1	20	1
		26	2

Note: One teacher had two qualifications, Dip Ed and B Sc (Comp)

One of the teachers from a double shift secondary school opted not to continue with the exercise after submitting his end of term 1 test (before peer instruction workshop). This resulted in 17 teachers, 2 females and 15 males, participating in the study up to the end. The 17 teachers were of varied qualifications and experience as shown in Table 3.2. As planned they were all teaching Physical Science in Form 3 in 2006 at selection and taught a 2007 Form 4 Physical Science class.

Selecting a teacher to participate in the study, was also selecting his or her class too for the study. The class played the role of examinees, a source of test scores. The study therefore involved 1543 learners as detailed later in section 3.5.1, Table 3.4.

3.3 School visits

During the study the schools were visited ten times, first in March 2007 to brief the participating schools about this study and arranging a research schedule. At the same time the selected teachers completed a questionnaire to show topics of interest to be included as content of the peer instruction workshop. The schools were visited thereafter four times to collect the tests for coding and again another four times to return learners' test scripts after

coding and also to pay the teachers for administering and marking the tests. Issues relating to the study were also discussed during such occasions. The last visit to the schools was at the end of the 2007 school year, to discuss with the participating teachers their perceptions of the exercise which took almost a full school year.

3.4 Teachers' peer instruction workshop in test construction

Teachers' peer instruction workshop lasted for four days. No instructor was involved. In the first two days teachers were discussing amongst themselves test construction ideas and procedures through group tasks given to them. In the last two days the teachers continued discussing and practicing the ideas and principles of test construction. The goal of teachers' peer instruction workshop was to enable the teachers to improve their test construction skills through active involvement (Lieb, 1991). Interaction with each other was enhanced through discussions about quality of test items and construction of classroom tests, in small groups mainly. In this way participants could have gained a deeper understanding of how classroom tests of high validity evidence could be constructed.

While at the workshop, each one of the participants was asked to construct, on their own, another test from the same test domain as their end of Term 1 test. The rationale for this was to find out whether or not the teachers would construct a test of higher validity evidence than the end of Term 1 test, after going through peer instruction in test construction.

The design of the teachers' peer instruction purposely did not provide for an instructor, as integrated ideas of peer instruction, constructivism and adult education seemed to suggest. Learning was based on participating teachers sharing with each other, through

discussions, their experiences and ideas about test construction. The arrangement was meant to create an environment similar to the one in schools, clusters and zones, Teacher Development Centres (TDCs), where teachers depend on each other's professional guidance. The study therefore intended to establish whether or not such guidance in test construction could be professionally effective.

A retired Physical Science teacher with a B.Ed. qualification was hired to provide neutral facilitation of the workshop. Facilitation was considered neutral in a sense that the facilitator was not an interested party in the outcome of the research which involved peer instruction workshop. His role was to control activities of the participants during the workshop, distribute tasks and resources to the groups, monitor that participants were carrying out their group tasks and assist with the evaluation of the workshop by distributing and collecting questionnaires. It was considered that this arrangement was going to ensure that the results of the workshop were not an influence of the researcher.

The facilitator was inducted on the workshop for a common understanding and approach to it with the researcher. It was emphasized that participants, on their own, were to find solutions for tasks given in the workshop based on their experience in assessment.

3.5 Data collection and instrumentation

Guided by research questions, the study focused on areas given in Table 3.3. As a result several instruments were constructed for data collection. The instruments were in three groups; teachers' tests, written questionnaires and unstructured discussion guide.

3.5.1 Teachers' pre-tests and post-tests

The main category of instruments for this study comprised tests constructed by teachers who formed the sample for the study. The tests were the focus for validity studies. See Appendix 3.2. In the process, teachers' performance in test construction before and after the peer instruction workshop in test construction was determined.

For the purpose of this study, each teacher was expected to construct and administer to a 2007 Form 4 Physical Science class four MSCE Physical Science tests; one end of Term 1 test (T1), another test from the same domain as end of Term 1 test (T2), 2007 mock test (M2) and any mock test (M1) previously administered. A total of 62 teachers' tests were collected for the study as shown in Table 3.4. T1 and M1 were constructed before teachers attended peer instruction while T2 and M2 were constructed after teachers attended peer instruction. Therefore, T1 and T2 formed one pre-test and post-test pair while M1 and M2 formed another. T1 and M1 were collected from the teachers on the second visit to the schools, which was during holidays of end of Term 1. M1 was given back to the schools for printing and administration when the mock tests were due in Third term. M1 was taken early in the study to ensure that schools would not know that it was going to be one of the tests until the printing time but to teachers only. Table 3.4 also shows details of the number of items from the tests, and the mean number of learners who wrote the tests.

Table 3.3 Focus area of the study

Research question	Data source	Type of information	Instrument
1. Were teachers' post-test items an equally relevant and representative sample of test domain as pre-test items?	1. Past mock tests and 2007 mock Physical science tests 2. End of term 1 test 3. Parallel test to end of term 1 test 4. Physical science Syllabus 5. Test blue print	1. Matching of items with content area	1. Item review form for raters
2. Did teachers' post-test items equally measure learners' corresponding cognitive ability levels as pre-test items?	1. Analysis results by SMEs	1. Proportion of number of items rated to have the highest average mean ratings for each order of cognitive level	1. Summary form of ratings of SMEs
3. Were the means of percentage total variances explained by common factors of teachers' post-tests equal to those of pre-tests?	1. EFA results	1. Percentage total variance explained by the tests	1. Factor analysis Output
4. To what extent were teachers aware of the need for raising validity evidence of their tests?	1. Teachers	1. Rating	1. Questionnaire
5. What were teachers' perceptions about possibilities of raising validity evidence of their tests through Peer instruction in test construction?	1. Teachers	1. Written responses 2. Verbal responses	1. Questionnaire 2. Unstructured discussion guide

Table 3.4 Number of tests, items and learners for the study

Teacher	No. of Tests	No. of items from the tests	Mean no. of learners that took the tests
1	4	175	117
2	2	82	83
3	4	190	52
4	4	193	65
5	4	207	125
7	4	243	72
8	4	228	65
9	2	125	159*
10	4	194	73
11	4	189	109
12	4	210	102
13	4	212	138
14	4	161	71
15	4	145	57
16	4	158	92
17	4	164	163
18	2	140	159*
Totals	62	3016	1543

Note: Teacher 9 and 18 were in the same school. Teacher 9 administered her two tests to the 159 learners when teacher 18 was not well. Teacher 18 administered his two tests to the same 159 learners because teacher 9 did not have M1 test.

Pre-tests and post-tests were planned to be administered within a two-week interval to reduce effects resulting from a long period between one administration and the other. For reasons beyond control in schools some were administered after two weeks. All these tests were theory papers. The practical component of M2 was not used because M1 had no practical component. The rationale was to compare impact of peer instruction on validity evidence in similar papers, theory papers. It was assumed that the pre-tests and post-tests were measuring MSCE Physical Science abilities of the learners.

The tests were administered to the whole class of Form 4 learners taught by a teacher. One school had a single stream while the rest of the schools had two or more streams. It was assumed that the learners were normally distributed in their classes in the schools. Table 3.5 shows the plan for administration of M1 and M2, and actual administration of the

tests in schools, which was intended to even out interfering variables due to administering of the tests at two different times. In the actual administration, two teachers did not have M1 tests while three other teachers did not follow instruction about which one of their tests was to start; M1 or M2.

Table 3.5 Administration of M1 and M2 in schools

	Planned		**Actual**	
Administration	M1	M2	M1	M2
First	8	9	5	10
Second	9	8	10	5

The tests as another category of instruments could not be pilot-tested or standardised in schools the way other instruments were treated to avoid leakage. Two of the submitted tests were used only in selected cases because they had some problems in other uses. One lacked mark allocation per item. Therefore, it could not be useful in cases where scores were the data of interest, in item analysis as an example. Another one needed to be re-coded but scripts had already been returned to schools. It also could not be used where scores were needed for analysis.

3.5.2 Questionnaires

The study also administered three written questionnaires. The first questionnaire was administered to prospective sample schools in the Southern Region of Malawi for baseline information. See Appendix 3.3. It solicited a teacher's personal information, qualifications and experience, among other things. The rationale for administering the questionnaire was

to identify teachers who qualified to be in the sample frame for the study. The questionnaire was given, for standardisation, to some officers at MANEB who had taught in secondary schools. MANEB has a large number of officers who were once teachers at some levels of the education system, i.e. primary, secondary, Teacher Training and Technical Colleges. Some items in the initial questionnaire were removed or modified and others added during standardization.

Another questionnaire was also administered to the sample to find out which test construction areas they wanted to be included in the content of teachers' peer instruction workshop. It was again standardized with officers at MANEB who once taught in schools and colleges. See Appendix 3.4. The rationale for administering it was to make content of peer instruction workshop more relevant, useful and motivating for the teachers (Conner, 2005).

The teachers were also required to show the extent to which they wanted the test construction areas included in the content of the workshop. After standardization of the questionnaire, the areas to be included in the content of the workshop were:

1. Definition of a classroom test
2. Description of a classroom test
3. Purposes of classroom tests
4. Domain of classroom tests
5. Characteristics of good test items
6. Order of test items
7. Test validity
8. Specification grid/test blue print for test construction

9. Marking schemes

10. Item analysis of classroom tests

A questionnaire was again administered to the teachers at the end of the workshop in order to evaluate the workshop. See Appendix 3.5. Like other questionnaires it was standardized with the help of an officer at MANEB. Workshop evaluation was necessary to gain an in-depth understanding of perceptions the teachers had about it.

3.5.3 Unstructured in-depth interview guide

At the end of the academic year in 2007, the study assessed teachers' experiences with test construction after going through the peer instruction workshop in order to establish their perceptions about peer instruction and test construction. Initially, a questionnaire and an unstructured in-depth interview guide were prepared for this purpose. During standardisation with officers at MANEB who had taught in secondary schools, it was recommended that the questionnaire be dropped. Items of the questionnaire would be part of the probing through in-depth interviews with the individual teachers. The predetermined inquiry areas for the in-depth interviews were usefulness of teachers' peer instruction workshop, its helpfulness, reasons teachers use past examination items in their tests, challenges the teachers encountered with application of test construction skills after the workshop and their recommendations. Focus group discussions would have been ideal for data collection. However, distance between the teachers was a limiting factor.

The researcher took notes during the in-depth interviews which lasted between two to two and half hours.

3.6 Data analysis

Several analyses were done during the study. They covered item analysis of teachers' tests, item relevance and representativeness rating for content related validity evidence, item cognitive representativeness and factor analysis for construct related validity evidence. Analyses were also done for teachers' perceptions about peer instruction and test construction before, during and after peer instruction workshop.

3.6.1 Item analysis of pre-tests and post-tests

After administering the tests, learners' marked scripts and copies of teachers' tests were collected for coding for item analysis. Table 3.4 shows the number of tests that were analysed in the study, including the number of items and learners who wrote the tests. The number of learners who wrote the tests translates into the number of scripts which were handled in the study.

Computed difficulty index, discrimination index and reliability coefficient served as indicators of quality in item analysis. The rationale for item analysis was to determine whether or not the pre-test and post-test pairs were similar with respect to test reliability, item difficulty and item discrimination. In addition, it was meant to guide test validation. A test needs to have at least moderate reliability in order to have some validity, and therefore to be considered for validation (Thorndike, 1997). In this regard, the recommended coefficients for moderate reliability range from 0.30 to 0.69 (Jackson, 2009).

P-p plot from SPSS was applied to check normality of distribution of data generated from computation of difficulty index, discrimination index and reliability coefficient. The rationale for it was to reinforce the assumption of normality of distribution and therefore to

determine possibility of applying t-test as described later in this section. Mann-Whitney U Test could have been used (Johnson, 1984), if p-p plot showed that the distribution was not normal.

Item difficulty

In this study the difficulty index (p_i) was calculated as given below:

$$p_i = \frac{\overline{X}}{X_{max}}$$

where

$\overline{X}$ is the mean score for candidates on the item and

X_{max} is the maximum score for the item

The rationale for using this formula was that items were polytomously scored**.** The acceptable range of neither too difficult nor easy items was 0.25 to 0.75 (Hopkins, 1998).

Paired sample t-test at $\alpha = 0.05$ was applied to determine significance of differences between means of number of items between a difficulty range in a pre-test and its corresponding difficulty range in the post-test. The result was going to be useful for interpreting such changes to the influence of peer instruction and influence of other factors between two administrations.

Item discrimination

Pearson product moment correlation was preferred because some of the items were polytomously scored (Crocker & Algina, 1986). As discussed in literature review, r is a

measure of how well an item discriminates between good examinees and poor ones. It was computed as,

$$r_{XY} = \frac{N(\sum XY) - (\sum X)(\sum Y)}{\sqrt{[N(\sum X^2) - (\sum X)^2][N(\sum Y^2) - (\sum Y)^2]}}$$

where,

N is the number of examinees attempting the item and

X , Y are the raw scores.

Paired sample t-test at $\alpha = 0.05$ was applied to determine the significance of differences between means of number of at least good items between pre-tests and post-tests. This was intended to find out whether or not there was going to be any significant change which could be due to peer instruction or other influences between one test administration and the other.

Test reliability

Reliability coefficient sought was alpha coefficient (α). The justification for using it was that the items were polytomously scored (Cronbach & Shavelson, 2004). It was computed using the equation below:

$$\alpha = \frac{k}{k-1}\left(1 - \frac{\sum \sigma_i^2}{\sigma_x^2}\right)$$

where

k is the number of items

$\sum \sigma_i^2$ is the sum of the variances of the items and

σ_x^2 the variance of the total score

Paired sample t-test at $\alpha = 0.05$ was applied to determine significance of differences between means of reliability coefficients between pre-tests and post-tests. The rationale was to determine if there was any significant change that could be due to peer instruction or effects of factors between one administration and the other.

Results of computation of item difficulty, item discrimination and test reliability using the equations given in this section are given in Tables 4.1 to 4.5 of Chapter 4 section 4.1.

3.6.2 Use of past examination items

An independent practicing Physical Science teacher was hired to go through teachers' tests to find out which of their items also appeared in past Malawi School Certificate of Education (MSCE) Physical Science examination papers from 1996 to 2006. Data generated was in the form of frequencies of observations made, which were converted into percentages. The rationale for looking for past examination items, was to establish their influence on the tests if they had been used and level of teachers' originality in test construction before and after peer instruction. P-p plot from SPSS was applied to test normality of distribution of percentages of past examination items in teachers' tests. Having found the distribution to be normal paired sample t-test at $\alpha = 0.05$ was applied to determine the significance of differences between means of number of past examination items in pre-tests and post-tests.

3.6.3 Content related validity evidence of pre-tests and post-tests

Pre-tests and post-tests were validated for content related validity evidence. In this process item relevance rating was done for all pre-tests and post-tests. However, item

representativeness was done only for T1 and T2. It could not be done for M1 and M2 tests because all M1 tests, except two, were theory papers only while M2 tests had both theory and practical papers, except two tests. The theory paper for M2 might have been at a numerical disadvantage generally compared with the theory paper for M1. The absence of test blue prints for M1 compounded the problem.

Six SMEs who were practicing Physical science teachers were involved to rate the relevance of the items to the content areas intended to be measured by the items in T1 and T2 tests. They were also markers of Junior Certificate of Education (JCE) and MSCE Physical Science examinations for MANEB. Orientation of the SMEs was done to ensure uniformity of rating procedures and accuracy of their results. The SMEs recorded their ratings on an item review form, adapted from Sireci (1998b).

SMEs were required to answer the following questions in the exercise:

1. To what extent did a given item measure given Physical Science topics of the test?
2. To what extent did a given item measure Physical Science abilities at given cognitive levels?

In order to answer question number 1; 'To what extent did a given item measure given Physical Science topics of the test?', the SMEs applied a relevance rating scale of 1 to 5. This was done in order to reduce subjectivity if a longer rating scale were used. The interpretation of this scale was that 1 meant an item measured that content area least while 5 meant an item measured that content area most. Applying the concept of Sireci (1998b) the index of item relevance was the mean relevance rating across the raters while an index of content area representation was the mean relevance index rating across all items

measuring the content area. In this study an acceptable relevant item rating was a minimum mean rating of 2.5 for an item on a topic. Number of relevant items across the topics was the criterion for determining representativeness of the items in the tests.

In order to justify the appropriateness of a t-test in determining item relevance, item representativeness in the test domain and item representativeness at cognitive levels, p-p plot using SPSS was applied to test normality of distribution of data generated from item relevance rating, item representativeness rating and item cognitive representativeness rating. Having established normality of the distribution of data, paired sample t-test was applied at $\alpha = 0.05$ to determine whether or not the differences between means of percentages of relevant items of the pre-test and post-test pairs were significant.

Item representativeness was compared at topic level between the tests. A bar graph of the frequencies of relevant items of topics in a test were plotted to compare whether or not T1 topics were more representative than T2 topics. The significance of the differences between means for both item relevance and item representativeness was tested using paired sample t-test at $\alpha = 0.05$ to determine whether or not peer instruction influenced the difference.

The same rating scale of 1 to 5 was used for question number 2, 'To what extent did a given item measure Physical Science abilities at given cognitive levels? A rating of 1 against an item on a cognitive level meant an item measured that level least while a rating of 5 against an item on a cognitive level meant an item measured the level most. Again an acceptable minimum mean rating of the cognitive level of an item was 2.5. The number of acceptable minimum mean ratings of cognitive levels of items was the criterion for determining whether or not a teacher's test was biased towards recall or comprehension or

higher order. In this regard, the pairs of the means of the frequencies, as percentages, an item tested a cognitive level were tested between the cognitive levels using paired sample t-test at $\alpha = 0.05$, to find out whether they were significantly different as an impact of peer instruction.

3.6.4 Construct related validity evidence of pre-tests and post-tests

Factor analysis statistical technique was applied to analyse construct related validity evidence of teachers' tests. The rationale for applying factor analysis was to compare influences of underlying common factors only, in terms of covariations in the pre-tests and post-tests variables (Tucker & McCallum, 1997). Therefore, CFA_1 was more suitable of the factor analysis techniques to apply in this case. The norm for comparison of construct related validity evidence between pre-tests and post-tests pairs was the percentage total variance the common factors explained. To determine the impact of peer instruction on construct validity evidence, paired sample t-test, at $\alpha = 0.05$ was used to test if there were any significant differences between the means of percentage total variance explained by the common factors of pre-tests and post-tests.

EFA model was applicable in the study because factor structure for the variables was not predetermined (Garson, 2007). CFA_2 would not have been appropriate for the study because it is used to test whether or not proposed results about specific subsets of variables actually define a factor as obtained in a previous research (Tucker &LaFleur, 1991). A different study applying a CFA_2 model would be required to verify the results of EFA of the current study.

Consideration was given to sample size for factor analysis in this study. Small sample sizes were involved in the study because class sizes were small too. Factor analysis in this case was guided by high communalities as much as was possible (MacCallum, et al. 1999; Hogarty, et al., 2005; Zhao, 2008). Communalities of at least 0.5 were recommendable (Field, 2005). In addition, Kaiser-Meyer-Olkin and Bartlett's Tests of Sphericity, and the determinant of the correlation matrix were applied to ensure appropriateness of conducting EFA (Pedhazur & Schmelkin, 1991). In some cases KMO of less than 0.5 was used depending on the strength of the other parameters. The same was true for communalities.

Principal axis factoring was preferred for factor extraction to maximum likelihood in that principal axis factoring is less likely to produce improper solutions compared to maximum likelihood. Besides this, the study did not intend to test for significance of factor loadings and correlations among factors (Fabrigar, et al., 1999 citing; Curran, West & Finch, 1996; Cudeck & O'Dell, 1994; Hu, Bentler & Kano, 1992). Besides this too, the solutions for factor extraction applying either principal factor analysis or maximum likelihood are very similar when normality is not severely violated (Fabrigar, et al. 1999).

For factor retention, eigenvalues greater than 1 was preferred in that it is simple and objective while the scree plot can also be subjective when the elbow is not clear (Fabrigar et al., 1999). Besides this the critical issue for conducting EFA in this study, was the proportion of variance attributable to the common factors, which is associated with eigenvalues.

Oblique rotation procedures were applied when more than one factor was extracted because they produce more accurate solutions than orthogonal rotations (Costello & Osborne, 2005; Fabrigar, et al., 1999). They are also more preferred because naturally

factors associated with psychological behaviour are most likely to be correlated (Rennie 1997, Conway & Huffcutt, 2003). In this regard promax procedure in oblique rotation was used because it is fast and allows replication by future studies while other oblique procedures do not (Garson, 2007; Abdi, 2003; Rennie 1997).

Like in content related validity evidence and item analysis, p-p plot from SPSS was applied to test normality of distribution of proportions of variance explained by common factors to find out the appropriateness of applying t-test in construct validation as well. When the results of p-p plot showed normality of the distribution, the differences between the means of the proportion of variance explained by common factors of the pre-tests and post-tests pairs were tested for significance applying paired sample t-test at $\alpha = 0.05$.

3.6.5 Analysis for teachers' perceptions about test construction

Simple mathematical procedures were applied to make sense out of data generated through questionnaires for determining test content and the questionnaire for evaluating the peer instruction workshop to tap teachers' perceptions about peer instruction. The data was manageable without complex statistical techniques or software like SPSS.

Content analysis was used for analyzing written notes, taken down during the in-depth interviews, about teachers' experiences with peer instruction and test construction (Denzin& Lincoln, 1998). Coding of written notes of the in-depth interview started soon after completing an interview with a teacher. The researcher took advantage of the many hours between one interview and the other because of the distance between the interviewees. During coding, a teacher's phrases describing experiences, in the written notes, were matched with those of other teachers. The objective was to categorise them,

initially, according to their similarities and differences within the predetermined areas of usefulness of teachers' peer instruction workshop, its helpfulness, reasons teachers use past examination items in their tests, challenges the teachers encountered with application of test construction skills after the workshop and their recommendations, a process known as open coding (Breg, 1997; Cresswell, 1994). Frequency tallies were used during categorization. After this the relationship of the categories was reassessed to determine main themes underlying the categorized descriptive narratives in a process referred to as axial coding (Narman, 1995). Selective coding was done in order to make conclusions about perceptions teachers had relating to peer instruction and test construction, and interpretations of the perceptions (Corbin, 1990).

3.7 Research ethics

The study was guided by three main ethical principles of respect for persons, beneficence and justice (APA, 2002; Department of Health, Education and Welfare, 1979). Respect for persons is demonstrated when participants are sufficiently informed and allowed to make voluntary decisions whether or not to participate in the study, and protection of those with diminished autonomy (Department of Health and Human Service, 2005). It includes personal privacy which applies to information gathered from the subjects (Malawi Government, 1999). Dissemination of such information needs their consent unless it is made anonymous. Beneficence is also recognized as an obligation of doing no harm to participants and maximizing their benefits as harm is being minimized (University of Washington School of Medicine, 2008). This harm can be physical, psychological, social and financial and rights based (Australian Government, 2008). Possible harm that could

have come from this type of research was psychological, social and financial and rights based. The fundamental issues of justice in research were noted to hinge on fair distribution of burdens and benefits, and rewards and penalty to subjects (Lebacqz, 1980).

3.7.1 Application of the principle of respect for persons

In order to comply with the principle of respect for persons the Heads of schools were written to seek information about the school and personal information of a 2006 Form 3 Physical Science teacher. See Appendix 3.8.The schools were made aware of the purpose of the baseline survey**.**

Through a letter and phone discussions, approval was sought from proprietors of schools to allow access to their schools that had the selected teachers. See Appendix3.9**.** The prospective schools and selected teachers were visited upon receipt of the approval. See Appendix 3.10. The visit provided the Heads of schools and the selected teachers with sufficient information about the proposed research exercise for their voluntary and informed consent to participate in the study (BERA, 2004). It also allowed Heads and teachers to ask questions on issues that needed further clarification to increase their level of understanding about what the research involved and what it meant to participate in the study and their right to withdraw on whatever ground (NASW, 1999; POST, 2008). In the process a schedule of research activities that was convenient to both the school and researcher was arranged.

Learners' scores from a teacher's tests were instrumental for the study. The researcher however, could not brief the learners directly about the purpose and procedures of the research to receive their individual voluntary and informed consent. It was left to the

teachers and the Heads of the schools to do it. Therefore, it was assumed that the learners had given voluntary consent for their participation based on informed decision.

3.7.2 Application of principle of beneficence

During the visit to schools perceived potential benefits of the exercise to schools, teachers, learners and the education system were discussed (APA, 2002). The teachers' voluntary consent to participate in the study was considered as an indicator that teachers perceived benefits from participating in the study.

3.7.3 Application of principle of justice

All the costs relating to participation in the study were met by the research budget (Department of Health, Education and Welfare, 1979). The teachers were also paid for administering and marking the tests. The rationale for the payment was to compensate them for a potential overload. The probability for them working odd hours was high because set schedules of the research activities were to be met.

Participating schools, teachers and learners were treated equally and sometimes according to demands of particular school's schedule (Lebacqz, 1980). A case in point was schools being at different stages of coverage of the MSCE Physical Science syllabus. As a result their schedules for mock tests were different. This had an effect on schedules for collecting learners' scripts for coding for item analysis and sending them back for use in preparation for MANEB examination which learners were preparing for. It had to be rearranged to suit individual school schedules at a cost to the researcher since schools did not have to be inconvenienced. At the same time affected research activities were adjusted

without having a serious adverse impact on the entire research schedule relating to the school, except increase in research expenses. These held up progress later as more money had to be sourced to complete the activities.

3.8 Limitations to the study

The proposed study, when implemented, experienced a number of unexpected issues. The issues might have affected the methodologies for carrying it out and time of completing it.

Sampling frame

The population of interest from which the sample of teachers was drawn was restrictive because of demographic factors. It was discovered that it had fewer Physical science teachers that satisfied the conditions. The sample was also unbalanced in terms of gender and type of schools from which they were selected for the same reason.

Literature source

Source of information was a big challenge. Most of the books in the Library at Chancellor College were old and few. The internet served as a supplementary source of information. However, internet service was very slow and sometimes it was not there for many days. In addition, some of the useful articles could not be accessed unless one was a member or bought the article. Considering time factor, it was not possible to purchase the articles on the internet or subscribe for membership.

Change in assessment practice

Teachers did not use test blueprints in their pre-tests. It might have been part of their assessment practice. This therefore meant that reference to test blueprints for pre-tests was not possible during the study. This affected in-depth analysis for comparison of content validation between pre-tests and post-tests.

Time constraint

The amount of time for the study, 3 – 4 years, was satisfactory. However, carrying out the study simultaneously with office work posed a big challenge. Sometimes it was a dilemma regarding whether or not attention should be given to the study or to the office demands. Consequently the study took longer to complete than envisaged.

3.9 Conclusion

The study was a mixed methods approach. It applied both quantitative and qualitative statistical techniques for data analysis. The rationale for it was to gain greater insight into the phenomenon of interest. The population of interest was small even for random sampling. As a result all the available subjects of interest had to be included in the sample.

The principles of ethics that governed conduct of the study were respect for persons, beneficence and justice. The subjects needed to know more about the study to make informed and voluntary decision of whether or not to participate, to benefit from participation and to be treated equally.

CHAPTER 4:

RESULTS AND DISCUSSIONS OF FINDINGS

4.0 Introduction

Chapter 4 presents the results and discussions of findings of the study. Item analysis has been presented focusing on item discrimination, item difficulty and test reliability. Test validation results, the core of the study, have been presented and discussed in the chapter. For this study, content and construct were sources of evidence for test validation since the tests are more for describing the learner than for decision making (Cronbach, 1971). Included in this Chapter are results and discussions of teachers' perceptions about peer instruction. The results have been generated through questionnaires and discussions with teachers. The purpose for administering the questionnaires and conducting interviews to the teachers was to have an idea of their perceptions about test construction before and after going through peer instruction, including application of test construction knowledge and skills during teaching at the time of this study. An overall summary has been presented to relate the impact on quality of teachers' items in terms of discrimination power and quality of their tests in terms of reliability and validity after peer instruction.

4.1 Item analysis of pre-tests and post-tests

Item analysis of the pre-test and post-tests was done. The equations for item discrimination, item difficulty and reliability given in section 3.6.1 of Chapter 3 have been applied to obtain results which are given in Tables 4.1 to 4.5 in this section. Item analysis is a preliminary requirement for test validation. It was done to determine which tests could be validated. The guiding factor was a minimum of moderate reliability (Thorndike, 1997), which ranges from 0.30 to 0.69 (Jackson, 2009). The analysis also gave an insight into whether or not the pre-test and post-test pairs were similar in terms of item discrimination, item difficulty and test reliability.

Item discrimination power, difficulty indices and reliability coefficients were computed as given in the following sections. P-p plot of distribution of discrimination indices, difficulty indices and reliability coefficients when applied was found to be normal. Therefore, t-test was applied to test for statistical significance of the differences between paired sample means at $\alpha = 0.05$ level. See Appendix 4.11

4.1.1 Item discrimination

Discrimination power is a measure of how effectively an item discriminates between good and poor examinees on the criterion of interest (Crocker & Algina, 1986). Crocker's and Algina's scale for the quality of items is given below:

Excellent	0.40 and above
Good	0.30 – 0.39
Mediocre	0.20 – 0.29
Poor	0.00 – 0.19

Worst below 0.00

Source: Crocker and Algina, (1986)

Quality of the pre-test and post-test items in this study, in terms of discrimination power, is summarised in Table 4.1 below as a percentage of the total items of a test.

Table 4.1 Percentage of good and excellent test items

Teacher	Percentage of good and excellent items combined in each form (D = 0.30 and above)			
	T1	T2	M1	M2
1	64	79	71	98
2	77	72	-	-
3	81	71	80	80
4	80	81	82	77
5	89	83	74	87
7	94	82	*	*
8	99	95	94	95
9	79	87	-	-
10	60	83	71	64
11	85	80	95	86
12	89	94	82	88
13	86	84	85	80
14	97	81	92	96
15	46	39	73	80
16	70	74	70	85
17	79	89	79	94
Mean	79.69	79.63	81.14	83.71

Key: T1 - end of term 1 tests (pre-tests), T2 - tests from the same domain as end of term 1 tests (post-test) M1 - past mock tests (pre-tests), M2 - 2007 mock tests (post-tests)
(-) - did not present one of the tests, (*) - one of the tests had no marks against items

Table 4.1 shows that after peer instruction, the mean of good and excellent items combined for T2 dropped by 0.06. However, the drop was not statistically significant (paired sample t-test, $p > 0.05$). See Appendix 4.12. There was no evidence to suggest that quality of teachers' items in T2 changed after peer instruction. T2 and T1 therefore were similar in terms of item quality.

The observation for mock tests was different. After peer instruction, the mean of good and excellent items combined increased by 2.57% for M2. However, like for T1 and T2, the difference between the means of good and excellent items combined for M1 and M2 was not statistically significant (paired sample t-test, $p > 0.05$). See Appendix 4.12. Again there was no strong evidence to claim that quality of teachers' items for M2 was different after peer instruction. M2 was therefore similar to M1 with respect to item quality.

Table 4.1 shows that generally the tests have a very high percentage of good and excellent items combined with the exception of T1 and T2 for Teacher 15. Quality of items of a test contributes to quality of a test but it is not a sufficient condition of quality of a test.

4.1.2 Item difficulty

Items were considered difficult in this study if their difficulty index (pi), was below 0.25, neither extremely difficult nor easy if the index was between 0.25 and 0.75 inclusive, and easy if the index was above 0.75 (Hopkins, 1998). Acceptable items for a test, therefore, had difficulty indices of the range between 0.25 and 0.75. Table 4.2 shows information relating to the three categories of difficulty levels of items used in T1 and T2 during this study.

Table 4.2 shows that after peer instruction the means of difficult, acceptable and easy items increased by 0.69%, dropped by 0.88% and increased by 0.19% respectively. The difficulty level among items of the three corresponding categories of difficulty levels of T1 and T2 did not change from one administration of the tests to the other. This is made evident by the difference between the pairs of the means in each of the given categories not being significant (difficult items: paired sample t-test, $p > 0.05$); acceptable items: paired

sample t-test, $p > 0.05$; easy items: paired sample t-test, $p > 0.05$). See Appendix 4.13. Therefore, it was considered that T2 was similar to T1 with respect to item difficulty.

Table 4.2 Percentage of T1 and T2 items at a given difficulty range

Teacher	T1			T2		
	P_i			P_i		
	<0.25	=>0.25 or <=0.75	>0.75	<0.25	=>0.25 or <=0.75	>0.75
1.	28	69	3	65	35	0
2.	19	68	13	21	65	14
3.	24	67	9	56	42	2
4.	77	20	3	40	60	0
5.	18	71	11	37	63	0
7.	56	44	0	69	31	0
8.	71	29	0	24	71	5
9.	35	63	2	50	50	0
10	50	50	0	64	36	0
11.	27	73	0	51	49	0
12.	37	61	2	28	60	12
13.	47	51	2	36	59	5
14.	81	19	0	55	43	2
15.	60	37	3	44	39	17
16.	53	44	3	59	41	0
17	54	41	5	49	49	2
Mean	46.06	50.44	3.5	46.75	49.56	3.69

Key: T1 - end of term 1 tests (pre-tests), T2 - tests from the same domain as end of term 1 tests (post-tests)

Difficulty indices for M1 and M2 are shown in Table 4.3. According to the information in Table 4.3 after peer instruction the means of difficult, acceptable and easy items of M2 increased by 3.69%, dropped by 1.15% and dropped by 2.53% respectively. However, the difference between the pairs of the means in each of the given categories was not significantly different except for easy items (difficult items: paired sample t-test, $p > 0.05$; acceptable items: paired sample t-test, $p > 0.05$; easy items: paired sample t-test, $p < 0.05$). See Appendix 4.13. The evidence which is there, only suggests a change in difficulty level of teachers' items between M1 and M2 in the category of easy items. Easier items were less

for M2 but this could not have affected other categories because of the small number of easy items. Again M2 was similar to M1 with respect to item difficulty.

Table 4.3 Percentage of M1 and M2 items at a given difficulty range

Teacher	M1			M2		
	P_i			P_i		
	<0.25	=> 0.25 or <=0.75	>0.75	<0.25	=> 0.25 or <=0.75	>0.75
1.	69	31	0	44	54	2
3.	48	46	6	53	47	0
4.	48	44	8	38	59	3
5.	33	62	5	27	67	6
8.	39	61	0	54	46	0
10.	37	54	9	51	47	2
11.	29	68	3	59	40	1
12.	21	65	14	10	74	16
13.	33	63	4	54	44	2
14.	28	64	8	55	45	0
15.	26	65	9	30	65	5
16.	52	46	2	42	58	0
17.	68	30	2	62	38	0
Mean	40.85	53.77	5.38	44.54	52.62	2.85

Key: M1 - past mock tests (pre-tests), M2 - 2007 mock tests (post-tests)

Low difficulty indices represent difficult items, which is an indicator of poor performance in a test. High difficulty indices from an item analysis reflect that items were easy. Nitko (1983) attributes poor performance in a test to poorly written items, incorrect prior learning and poor motivation for tests. Applying Nitko's observation, since the difference between means of difficult items and acceptable items for M1 and M2 was not significantly different, quality of items with respect to item difficulty, and learners' preparations for pre-tests and post-tests did not differ much.

The percentage of easier items dropped in the category of easy items for M2 compared to M1. It could not be due to learners' poor preparation and low motivation for the test because first the administration had 5 of M1 and 10 of M2 while the second administration

had 10 of M1 and 5 of M2 to equate groups for the influence of extraneous variables. The observed reduction in percentage of easy items in M2 should be an effect of peer instruction. Nonetheless, it did not have an effect on the other categories because of the small numbers of items involved in the category of easy items.

4.1.3 Reliability of pre-tests and post-tests

Alpha reliability coefficients were also determined for pre-tests and post-tests. Table 4.4 shows alpha reliability coefficients for T1 and T2. It was observed from Table 4.4 that the mean of alpha reliability coefficients for T2 dropped by 0.0008. The difference between the means however, was not statistically significant (paired sample t-test, $p > 0.05$). See Appendix 4.12. This suggests that T1 and T2 were similar with respect to test reliability.

Table 4.5 shows alpha reliability coefficients for M1 and M2. The mean alpha reliability coefficient for M2 dropped by 0.0123.

The difference between the means however, was not statistically significant (paired sample t-test, $p > 0.05$). See Appendix 4.12. Again there was no statistical evidence to support any claim that reliability coefficients changed for M2. M2 was similar to M1 with respect to test reliability.

Referring to information in Table 4.4 and Table 4.5, it is observed that all of the tests generally had high reliability coefficients. However, as discussed in Chapter 2, reliability is not a sufficient condition for quality of a test.

Table 4.4 Alpha reliability coefficients of T1 and T2

	Alpha coefficient	
Teacher	T1	T2
1	0.8861	0.7631
2	0.9000	0.8302
3	0.8881	0.8177
4	0.8232	0.9192
5	0.9346	0.9205
7	0.9616	0.9194
8	0.9063	0.9400
9	0.8705	0.9444
10	0.9051	0.8601
11	0.9001	0.9459
12	0.9158	0.8711
13	0.8354	0.8299
14	0.9390	0.9436
15	0.8403	0.8410
16	0.7865	0.9096
17	0.9005	0.9248
Mean	0.8871	0.8863

Key: T1 - end of term 1 tests (pre-tests), T2 - tests from the same domain as end of term 1tests (post-tests)

Table 4.5 Alpha reliability coefficient of M1 and M2

	Alpha coefficient	
Teacher	M1	M2
1	0.7619	0.8476
3	0.8447	0.6238
4	0.8714	0.8168
5	0.8239	0.8318
8	0.8847	0.8996
10	0.8991	0.9070
11	0.8204	0.8725
12	0.8752	0.8321
13	0.8165	0.8053
14	0.9414	0.9163
15	0.9564	0.8909
16	0.9040	0.9209
17	0.8372	0.9127
Mean	0.8644	0.8521

Key: M1 – past mock tests (pre-tests), M2 – 2007 mock tests (post-tests)

4.1.4 Summary: Item analysis results

The results of item analysis show that item discrimination, item difficulty and reliability of the pre-tests and post-tests were generally the same. This was an indicator that learners took similar tests with respect to item discrimination, item difficulty and test reliability between the first and the second test administration. Quality of items with respect to item discrimination was generally high. Therefore teachers were able to identify for their tests good and excellent test items before and after peer instruction, which was an indicator that they had some knowledge of qualities of good items.

Alpha reliability coefficients for the tests ranged from 0.6238 to 0.9616. Test reliability therefore ranged from moderate to high (Jackson, 2009). Since reliability is a pre-requisite for validity and that tests must have at least moderate reliability to have some validity, all the tests given in Tables 4.4 and 4.5 were suitable for test validation study (Oosterhof, 2001; Thorndike, 1997).

The other thing to learn from the results of item analysis as reflected by discrimination power and difficulty indices of items as well as the alpha reliability coefficient of the tests is that learners lacked mastery of Physical Science subject matter. The alpha reliability coefficient should have been much lower than observed to reflect masterly of subject matter. The items too would have been less discriminating. This is as long as the difficulty indices for the items were greater than 0.75 and also not less than 0.25. A difficulty index of less than 0.25 is an indicator that learners were performing poorly on those items, i.e. items were difficult while an index of greater than 0.75 means that learners were scoring well in those items; i. e. items were easy (Hopkins, 1998).

The teachers participating in the study were of varying ages, qualification and experience. They were also trained in different colleges and were teaching in different

schools. The results of item analysis do not show a trend that could be attributed to the different characteristics of the teachers. May be the trend might show with larger samples. In this regard, there is no evidence to claim that teacher performance in test construction was dependent on their age, qualification, experience, college they attended and school environment in which they were teaching.

4.2 Content related validity evidence of pre-tests and post-tests

Two of the questions the study sought to answer were 'Were teachers' post-test items an equally relevant and representative sample of test domain?' and 'Did teachers' post-test items equally measure learners' cognitive ability levels as pre-test items?' To answer the questions, pre-tests and post-tests were analysed for item relevance and item representativeness, in addition to item cognitive representativeness. The results of the analysis assisted to establish whether or not content related validity evidence increased with peer instruction, besides teachers' practice of testing cognitive levels. The results of the analysis are given in this section. P-p plot for distribution of data for item relevance, item representativeness and item cognitive representativeness was found to be normal. On this account, paired sample t-test was applied to test if there was a statistically significant difference between sample means at $\alpha = 0.05$. See, Appendix 4.1.

4.2.1 Item relevance rating of pre-tests and post-tests

Table 4.6 shows results of relevance rating of T1 and T2 as a percentage of frequencies of relevance rating of test items. See Appendix 4.15**.** The information in Table 4.6 shows that T1 had more relevant items than T2. However, statistically this was not significant (paired sample t-test, $p > 0.05$). See Appendix 4.16. Therefore, there was no sufficient evidence to suggest that item relevance rating changed through peer instruction for T2.

Table 4.6 Percentage of relevant items of T1 and T2

Teacher	T1			T2		
	Number of items of a test	Number of relevant items	Percentage (Relevance)	Number of items of a test	Number of relevant items	Percentage (Relevance)
1	32	32	100	99	97	97.98
2	42	41	97.62	33	33	100
3	32	32	100	43	43	100
4	51	51	100	45	45	100
5	45	45	100	61	60	98.36
7	54	54	100	72	72	100
8	61	61	100	43	42	97.70
9	52	50	96.15	54	54	100
10	33	33	100	38	38	100
11	33	33	100	44	44	100
12	59	59	100	49	48	97.96
13	57	52	91.23	60	56	93.33
14	38	38	100	63	61	96.83
16	34	34	100	40	40	100
17	37	37	100	43	43	100
Mean			99			98.81

Key: T1 - end of term 1 tests (pre-tests), T2 - tests from the same domain as end of term 1tests (post-tests)

Table 4.7 shows the results of relevance rating for M1 and M2. As observed from Table 4.7, the general trend was that mean item relevance rating for M2 was less than mean item relevance rating of M1 but this was not statistically significant paired sample t-test, $p > 0.05$). See Appendix 4.16. This implied that peer instruction had no impact on raising

item relevance for M2. It seems obvious since teachers' tests are supposed to test objectives laid down against a topic in a syllabus. In such cases chances are slim to stray off the domain of interest during test construction.

Table 4.7 Percentage of relevant items of M1 and M2

Teacher	M1			M2		
	No. of items of a test	No. of relevant items	Percentage (Relevance)	No. of items of a test	No. of relevant items	Percentage (Relevance)
1	70	57	81.43	54	53	98.15
3	72	60	83.33	43	42	97.67
4	67	67	100	55	55	100
5	59	59	100	53	52	98.11
7	64	64	100	63	63	100
8	83	71	85.54	67	66	98.51
10	54	54	100	75	75	100
11	67	59	88.06	84	72	85.71
12	83	81	97.59	55	51	92.73
13	67	67	100	52	49	94.23
14	34	34	100	67	67	100
15	43	43	100	34	34	100
16	41	29	70.73	34	34	100
17	68	61	89.71	40	40	100
18	73	73	100	67	67	100
Mean			93.09			97.67

Key: M1 - past mock tests (pre-tests), M2 - 2007 mock tests (post-tests)

4.2.2 Rating for representativeness of items of T1 and T2

Items of T1 and T2 were analysed for representativeness. It should be noted that items for T1 and T2 were from the same test domain. Item representativeness was not determined for M1 and M2 for reasons given earlier. Table 4.8 is a summary of topic-by-topic comparison of item representativeness between T1 and T2.

Table 4.8 Number of topics with better item representation

Teacher	T1	T2
1	1	5
2	3	1
3	0	2
4	1	4
5	1	5
7	1	5
8	3	2
9	3	3
10	3	2
11	0	1
12	3	3
13	3	2
14	0	6
15	2	2
16	1	3
17	2	5
Mean	1.69	3.19

Key: T1 - end of term 1 tests (pre-tests), T2 - tests from the same content domain as end of term 1tests (post-tests)

The data in Table 4.8 is also presented graphically in Figure 4.1 and Figure 4.2 as samples to show how item representativeness compares between pre-tests and post-tests at topic level for Teacher 1 and Teacher 2 respectively. See Appendix 4.18 for more graphs of all the teachers.

Graphical presentation of item representativeness for T1 and T2 for Teacher 1 and Teacher 2 shows the same result as described above for Teacher 1 and Teacher 2 and their tests as given in Table 4.8.

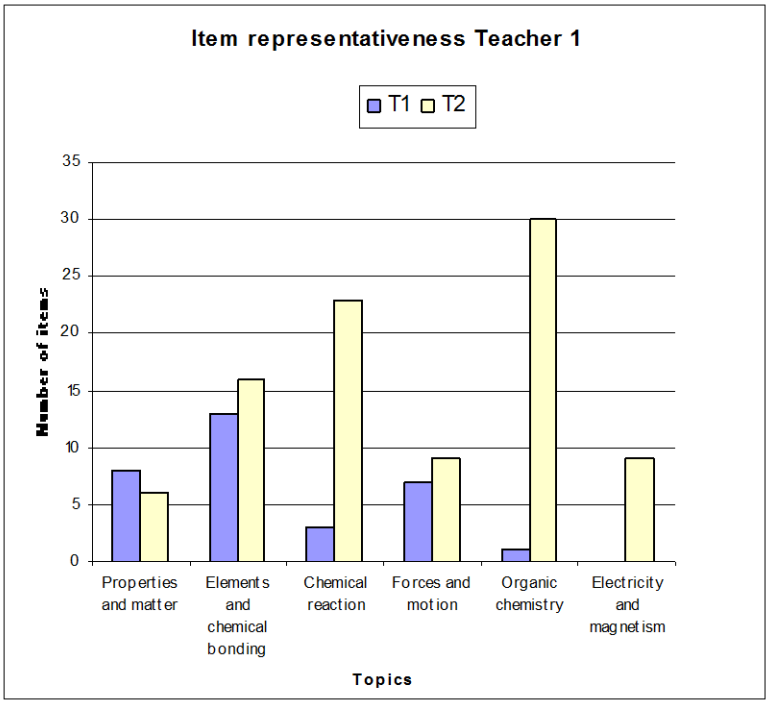


Figure 4.1 Item representativeness at topic level: Teacher 1

The trend for all the tests, as observed from Table 4.8, was that items for T2 topics were more representative than items of T1 topics. Therefore, items for T2 tests were more representative than items for T1 tests. The observation was supported statistically, since the difference between the means of more item representative topics of T1 and T2 was significant (paired sample t-test, $p < 0.05$) in favour of T2. See Appendix 4.16.

It was confirmed further from both Table 4.8 and graphs in Appendix 4.18 as 56.25% of T2 had more topics with a better item representation than T1 while 25% of T2 had less than T1. Another 18.75% of T2 had the same number of topics with a better item representation in the topic as T1. The proportions too support the claim that content validity evidence of T2 increased as an impact of peer instruction.

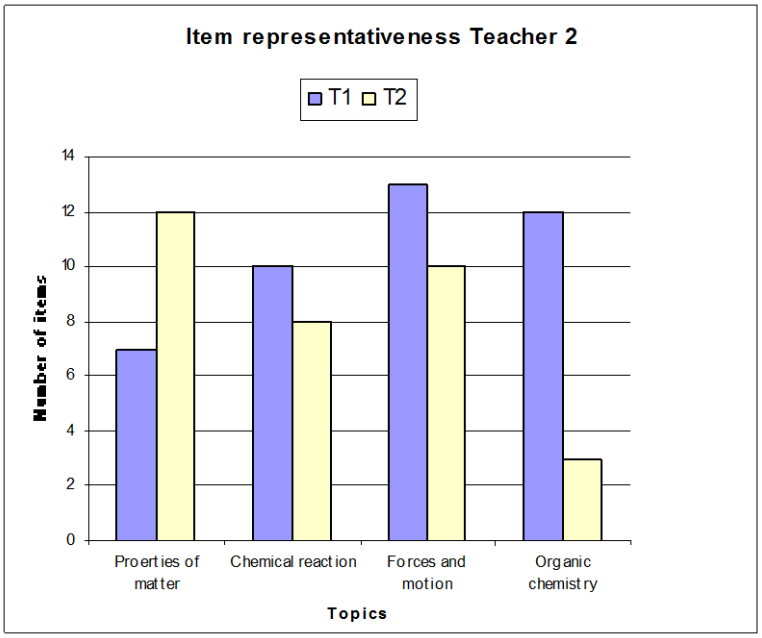


Figure 4.2 Item representativeness at topic level: Teacher 2

4.2.3 Cognitive rating of items for teachers' pre-tests and post-tests

Teachers' test items were analysed to determine the cognitive level at which the items tested given topics. Table 4.9 and Table 4.10 show the results of the analysis in terms of percentage of T1, T2, M1 and M2 items at each cognitive level. The analysis is based on items which were rated relevant for the level. Since it involved rating by SMEs some items were rated to test a topic at multiple cognitive levels while others showed to test no level at all at 2.5 as a minimum relevance mean rating. P-p plot was applied to check normality of distribution of percentages of teachers' items at each cognitive level. In this regard, paired sample t-test at $\alpha = 0.05$ was applied to test if the difference between paired sample means between cognitive levels were statistically significant.

Table 4.9 Percentage of T1 and T2 items at a cognitive level

Teacher	T1			T2		
	R	C	H	R	C	H
1	23.08	56.41	20.51	35.29	41.18	23.53
2	35.94	59.30	13.95	21.62	37.84	40.54
3	48.49	15.15	36.36	40.74	33.33	25.93
4	34.00	42.00	24.00	42.86	28.57	28.57
5	36.84	35.09	28.07	34.33	41.79	23.88
7	44.44	28.57	26.98	29.63	41.98	28.40
8	30.67	46.67	22.67	25.58	58.14	16.28
9	29.21	51.69	19.10	29.51	47.54	22.95
10	40.91	34.09	25.00	30.43	43.48	26.09
11	28.95	42.11	28.95	41.86	37.21	20.93
12	47-69	32.31	20.00	47.27	30.91	21.82
13	30.51	44.07	25.42	39.19	47.30	13.51
14	45.00	37.50	17.50	43.86	35.09	21.05
15	32.50	32.50	35.00	43.75	25.00	31.25
16	17.39	39.13	43.48	35.39	35.29	29.41
17	23.26	34.88	41.86	27.27	13.64	59.09
Mean	29.73	38.52	27.14	35.34	37.67	26.95

Key: T1 - end of term 1 tests (pre-tests), T2 - tests from the same domain as end of term 1tests (pre-tests) R – recall, C – comprehension, H – higher order

Table 4.10 Percentage of M1 and M2 items at a cognitive level

Teacher	M1			M2		
	R	C	H	R	C	H
1	19.12	57.35	23.53	27.94	35.29	36.76
3	26.67	54.67	18.67	35.59	37.29	27.12
4	30.43	40.58	28.99	29.41	35.29	35.29
5	42.22	26.67	31.11	26.47	48.53	25.00
7	46.67	28.00	25.33	30.56	44.44	25.00
8	32.14	38.10	29.76	33.33	41.98	24.69
10	48.39	25.81	25.81	41.11	41.11	17.78
11	34.78	46.38	18.84	49.04	31.73	19.23
12	63.64	27.27	9.09	51.67	33.33	15.00
13	55.26	26.32	18.42	30.51	47.46	22.03
14	30.95	50.00	19.05	39.24	32.91	27.85
15	57.45	23.40	19.15	40.00	35.00	25.00
16	24.19	45.16	30.65	18.52	37.04	44.44
17	29.89	45.98	24.14	35.85	35.85	28.30
18	41.18	35.29	23.53	37.50	43.18	19.32
Mean	38.87	38.07	23.07	35.12	38.70	26.19

Key: M1 – past mock test (pre-tests), M2 - 2007 mock test (pre-tests)
R – recall C – comprehension H – higher order

The means between corresponding cognitive levels for the pre-tests and post-tests were tested for significant difference using t-test. Table 4.11 below shows the computed p-values of the t-test at $\alpha = 0.05$. Since the computed p-values shown in Table 4.11 are greater than $p = 0.05$ it means that statistically the means are not different. Teachers' practice of testing lower order cognitive levels more than higher order level was the same before and after peer instruction. The interpretation of this was that peer instruction did not change teachers' practice of testing lower cognitive levels more than higher order cognitive levels. See, Appendix 4.17.

Table 4.11 P-values for differences between means ($\alpha = 0.05$)

	Pre-test and post-test pairs	
Corresponding cognitive level	T1/T2	M1/M2
Recall	0.649	0.236
Comprehension	0.754	0.873
Higher order	0.952	0.099

Table 4.12 also shows computed p-values of t-test at $\alpha = 0.05$ to test significant difference between the means of cognitive levels within a set of the pre-tests and post-tests.

From Table 4.12 the computed p-values for the difference between means of comprehension and higher order in T1, recall and higher order, and comprehension and higher order in T2, M1 and M2 are less than $p = 0.05$. This means that the difference between the means of comprehension and higher order in T1, recall and higher order, and comprehension and higher order in T2, M1 and M2 are significantly different in favour of a lower order since the lower order has a larger mean. See appendix 4.17. The interpretation is the same that generally the tendency for teachers was to test lower order cognitive levels more than higher order before and after attending peer instruction in test construction.

Table 4.12 P-values for differences between means (α = 0.05)

		Recall	Comprehension	Higher order
T1	Recall			
	Comprehension	0.315		
	Higher order	0.065	0.009	
		Recall	Comprehension	Higher order
T2	Recall			
	Comprehension	0.546		
	Higher order	0.043	0.042	
		Recall	Comprehension	Higher order
M1	Recall			
	Comprehension	0.899		
	Higher order	0.003	0.000	
		Recall	Comprehension	Higher order
M2	Recall			
	Comprehension	0.267		
	Higher order	0.042	0.000	

4.2.4 Summary: Content related validity evidence

Evidence was not sufficient to claim that peer instruction had an impact on item relevance of post-tests. In terms of item representativeness, there was sufficient evidence statistically to claim that peer instruction had an impact on item representativeness for T2. Therefore, the claim that content related validity evidence, with respect to item representativeness, for T2 increased through peer instruction in test construction was supported. Content related validation for M1 and M2 was not done for reasons given earlier. The results of the analysis also did not have evidence to support any claim that peer instruction had an impact on teachers' practice of testing the cognitive levels of the post-tests. Teachers' tendency was for more recall and comprehension items than higher order items in both pre-tests and post-tests. This confirms the concerns that classroom test items are mostly low order.

Like in item analysis there was no trend in the results of item relevance rating, item representative rating and item cognitive rating that could be attributed to the influence of the teachers' age, qualification, experience, college where they were trained and school environment where they were teaching.

4.3 Construct related validity evidence of pre-tests and post-tests

EFA was applied in order to answer the question 'Were the means of percentage total variances explained by common factors in the teachers' pre-tests and post-tests the same?' See appendix 4.19 for sample of EFA analysis. The intention was to establish whether or not construct validity evidence of the post-tests increased as an effect of peer instruction. The criterion measure for supporting that construct validity evidence increased or not, was the proportion of the variance attributed to common factors. The results of the analysis are given in the sub-sections which follow. The result of p-p plot for distribution of percentage variance explained by common factors showed that the data was normally distributed. Therefore, paired sample t-test could be applied to test for statistical significance of the difference between paired sample means in Tables 4.11 – 4.14. See Appendix 4.11

4.3.1 EFA of T1 and T2 at question level

EFA of T1 and T2 was done at question level. Table 4.13 shows the summary of the results of the analysis. Using the parameters for data quality discussed in Chapter 3, it was observed that all T1 and T2 given in Table 4.13 were appropriate for EFA. The observed determinant values of 0.000 in Table 4.13 for some of the tests do not imply that they were

less than 0.00001. It is only that SPSS gives the values to three decimal places otherwise it was not going to be possible to find their factor solutions.

Table 4.13 EFA results for T1 and T2 at question level

Teacher	T1					T2				
	Det	KMO	BTS	No. CF	% TVE	Det	KMO	BTS	No. CF	% TVE
1	0.450	0.645	0.000	2	44	0.019	0.874	0.000	1	52
2	0.041	0.880	0.000	1	51	0.225	0.789	0.000	1	55
3	0.025	0.852	0.000	1	43	0.073	0.601	0.000	2	47
4	0.000	0.819	0.000	2	56	0.000	0.864	0.000	4	61
5	0.012	0.916	0.000	1	61	0.008	0.929	0.000	1	64
7	0.003	0.861	0.000	2	58	0.001	0.860	0.000	1	59
8	0.004	0.828	0.000	1	64	0.008	0.885	0.000	1	62
9	0.008	0.917	0.000	1	57	0.020	0.866	0.000	1	55
10	0.253	0.720	0.000	2	46	0.033	0.824	0.000	1	50
11	0.289	0.703	0.000	1	40	0.311	0.729	0.000	1	48
12	0.002	0.919	0.000	1	54	0.000	0.940	0.000	1	67
13	0.006	0.917	0.000	1	59	0.004	0.931	0.000	1	63
14	0.039	0.853	0.000	1	50	0.028	0.900	0.000	1	61
15	0.233	0.754	0.000	2	41	0.256	0.786	0.000	1	43
16	0.065	0.841	0.000	1	50	0.020	0.843	0.000	3	51
17	0.041	0.901	0.000	1	51	0.032	0.900	0.000	1	53
Mean					52					56

Key: Det. – determinant, KMO - Kaiser – Meyer – Olkin, BTS - Bartlett's Test of Spherecity, CF - common factors, %TVE – percentage total variance explained (proportion of variance attributed to common factors), T1 - end of term 1 tests (pre-tests), T2 - tests from the same domain end of term 1 tests (post-tests)

Table 4.13 shows that after peer instruction the mean proportion of variance attributed to common factors in T2 was 56% and 50% for T1. The difference between the means was significant (paired sample t-test, $p < 0.05$). See Appendix 4.20. The interpretation therefore, was that construct validity evidence of T2 increased through peer instruction. Considering individual tests in Table 4.13, 13 out of 16 T2 representing 81.25 % had an increase in percentage total variance explained by the tests. It can therefore be said 81.25% of the T2 showed an increase of construct validity evidence. The high percentage of T2 with an increase of percentage total variance explained by common factors confirms the significance of the difference between the means.

4.3.2 EFA of M1 and M2 at question level

EFA was also conducted for M1 and M2 at question level. Table 4.14 shows the summary of the results of the analysis. The observed 0.000 values of determinant in Table 4.13 have the same explanation as given for observation made in Table 4.13.

M1 and M2 were also appropriate for EFA with respect to data quality and reliability. M1 for school 4 has a KMO value less than 0.5. Its suitability for EFA is based on its reliability, determinant value and Bartlett's test of Sphericity tests.

Table 4.14 shows that after peer instruction the mean proportion of variance attributed to common factors for M2 is 60.86% and 56.29% for M1. The difference between the means was significant (paired sample t-test, $p < 0.05$). See Appendix 4.20. Again, information was adequate to claim that construct related validity evidence of M2 increased

Table 4.14 EFA results for M1 and M2 at question level

Teacher	M1					M2				
	Det	KMO	BTS	No. CF	% TVE	Det	KMO	BTS	No. CF	% TVE
1	0.019	0.901	0.000	1	57	0.001	0.947	0.000	1	66
3	0.000	0.804	0.000	1	53	0.022	0.865	0.000	1	48
4	0.001	0.886	0.000	1	54	0.000	0.894	0.000	4	66
5	0.005	0.922	0.000	1	60	0.001	0.917	0.000	1	65
8	0.008	0.877	0.000	1	67	0.000	0.913	0.000	1	75
10	0.074	0.854	0.000	1	39	0.009	0.861	0.000	1	53
11	0.004	0.920	0.000	1	62	0.001	0.929	0.000	1	68
12	0.007	0.910	0.000	1	64	0.007	0.911	0.000	1	64
13	0.002	0.883	0.000	1	63	0.009	0.922	0.000	1	57
14	0.021	0.786	0.000	1	55	0.001	0.944	0.000	1	71
15	0.046	0.847	0.000	1	49	0.026	0.856	0.000	1	52
16	0.007	0.881	0.000	1	46	0.012	0.885	0.000	1	54
17	0.004	0.874	0.000	1	51	0.014	0.918	0.000	1	55
Mean					55					61

Key: Det. – determinant, BTS - Bartlett's Test of Spherecity, KMO - Kaiser – Meyer – Olkin
CF -common factors, %TVE – percentage total variance explained (proportion of variance attributed to common factors), M1 - past mock tests (pre-tests), M2 - 2007 mock tests (post-tests)

as an influence of peer instruction. From a different perspective, an assessment of individual tests shows that 9 out of 14 M2, representing 64.29 % of M2 had an increase in the proportion of variance attributed to common factors. Therefore 64.29% of M2 showed an increase of construct related validity evidence.

The high percentage of tests having an increase in percentage total variance explained by common factors reinforces the interpretation that peer instruction had a positive impact on construct validity evidence of M2.

4.3.3 EFA of T1 and T2 at sub-question level

Interpretation of 'sub-question' was that if a question, for example question 1, has 1a and 1b, 1a and 1b were sub-questions.

EFA of T1 and T2 was also carried out at sub-question level. Table 4.15 shows the summary of results of the analysis. See Appendix 4.20 for detailed results of EFA at sub-question level. T2 for school 3 has a KMO value less than 0.5. Its suitability for EFA was based on the other conditions in addition to the fact that factor extraction was not terminated.

Table 4.15 shows that after peer instruction the mean proportion of total percentage of variance explained by T2 was 54.56% and 52.63% for T1. The difference between the means was not statistically significant (paired sample t-test, $p > 0.05$). See Appendix 4.20.

Individual tests show that 9 out of 16 T2 representing 56.25 % of T2 had an increase in the percentage total variance explained by the tests. Therefore 56.25% of T2 show an increase of construct validity evidence. This was too small a proportion to show a significant difference between the means of T1 and T2.

Table 4.15 EFA results for T1 and T2 at sub-question level

Teacher	T1					T2				
	Det	KMO	BTS	No. CF	% TVE	Det	KMO	BTS	No. CF	% TVE
1	0.080	0.698	0.000	4	39	0.000	0.761	0.000	12	59
2	0.000	0.765	0.000	6	52	0.010	0.765	0.000	5	43
3	0.000	0.603	0.000	9	58	0.000	0.392	0.000	7	69
4	0.000	0.665	0.000	10	70	0.000	0.628	0.000	10	63
5	0.000	0.901	0.000	6	49	0.000	0.906	0.000	5	47
7	0.000	0.762	0.000	7	56	0.000	0.705	0.000	8	63
8	0.000	0.820	0.000	4	65	0.000	0.816	0.000	5	59
9	0.000	0.695	0.000	11	59	0.000	0.867	0.000	6	53
10	0.002	0.658	0.000	7	47	0.000	0.767	0.000	7	55
11	0.030	0.572	0.000	6	54	0.010	0.818	0.000	4	45
12	0.000	0.700	0.000	8	58	0.000	0.907	0.000	7	58
13	0.000	0.887	0.000	6	51	0.000	0.883	0.000	7	51
14	0.000	0.829	0.000	4	46	0.000	0.843	0.000	7	59
15	0.036	0.723	0.000	4	45	0.070	0.533	0.000	3	45
16	0.000	0.717	0.000	8	50	0.000	0.687	0.000	9	54
17	0.000	0.848	0.000	7	44	0.000	0.874	0.000	6	43
Mean					53					54

Key: Det. – determinant, BTS - Bartlett's Test of Spherecity, KMO - Kaiser – Meyer – Olkin CF - common factors, %TVE – percentage total variance explained (proportion of variance attributed to common factors), T1 - end of term 1 tests (pre-tests), T2 - tests from the same domain as end of term 1tests (post-tests)

It could be said that peer instruction had a negligible impact on construct validity evidence of the tests at sub-question level. The only argument is that the tests that showed an increase of construct related validity evidence at question level could have little or no construct related validity evidence when tested at sub-question level.

4.3.4 EFA of M1 and M2 at sub-question level

EFA of M1 and M2 was also done at sub-question level**.** Table 4.16 shows the summary of the results of the analysis.

Considering information in Table 4.16, the mean percentage of total percentage of variance explained by common factor for M2 was 55.85% and 56.15% for M1. The

difference between the means was not statistically significant (paired sample t-test, $p > 0.05$). See Appendix 4.20.

Therefore there was no sufficient evidence to claim that peer instruction had an impact on construct related validity evidence of M2 at sub-question level. The same observation about different construct related validity evidence at question and sub-question level applies for M2 like in T1 and T2. Analysis of individual tests shows that 4 out of 13 M2 representing 30.77 % of M2 had an increase in the percentage total variance explained by the common factors. Therefore 30.77% of M2 show an increase of construct validity evidence. The percentage of tests showing an increase of percentage total variance explained by common factors was low. It was reasonable to claim that peer instruction had negligible impact on raising construct validity evidence of M2.

Table 4.16 EFA results for M1 and M2 at sub-question level

Teacher	M1					M2				
	Det	KMO	BTS	No. CF	% TVE	Det	KMO	BTS	No. CF	% TVE
1	0.000	0.841	0.000	7	48	0.000	0.943	0.000	3	53
3	0.000	0.706	0.000	8	66	0.000	0.706	0.000	4	57
5	0.000	0.905	0.000	5	49	0.000	0.902	0.000	6	52
8	0.000	0.897	0.000	3	58	0.000	0.696	0.000	8	75
10	0.000	0.600	0.000	9	54	0.000	0.264	0.000	13	68
11	0.000	0.841	0.000	5	54	0.000	0.905	0.000	6	58
12	0.000	0.796	0.000	6	56	0.000	0.808	0.000	7	56
13	0.000	0.897	0.000	6	54	0.000	0.895	0.000	5	49
14	0.000	0.738	0.000	4	63	0.000	0.796	0.000	7	60
15	0.000	0.784	0.000	4	53	0.002	0.776	0.000	4	47
16	0.000	0.731	0.000	11	53	0.000	0.862	0.000	6	53
17	0.000	0.811	0.000	13	56	0.001	0.922	0.000	3	45
Mean					55					56

Key: Det. – determinant, KMO - Kaiser – Meyer – Olkin, BTS - Bartlett's Test of Spherecity, CF - Common Factor, %TVE – percentage total variance explained (proportion of variance attributed to common factors), M1 - past mock tests (pre-tests), M2 - 2007 mock tests (post-tests)

4.3.5 Summary: Construct related validity evidence

Construct related validity evidence increased for T2 and M2 at question level. In both cases it was statistically significant, while at sub-question level the increase was not statistically significant. See Appendix 4.9. There was enough evidence to attribute the increase of construct validity evidence to peer instruction at question level.

At sub-question level information was not sufficient to support the claim that peer instruction had an impact on construct validity evidence. The contrasting result of EFA at sub-question level from that at question level was an issue of interest. Sub-question level exploratory factor analysis was based on data that had been rearranged from question level data. The rearrangement might have resulted in the data being unsuitable for plausible factor extraction. Inter-dependence between sub-questions might have been increased as a result the sub-questions became highly correlated. At question level, the questions were likely to be independent of each other. Consequently, inter-correlation between them was just adequate for successful factor extraction.

The observation might also be an indicator of some characteristic of factor analysis. In the rearranged data exploratory factor analysis might be measuring different traits because the abilities might also have been rearranged. Hoste (1982) had a surprise result of his exploratory factor analysis of Biology theory and practical examination. Instead of loading on different factors the two examinations loaded on the same factor with theory loading more highly on it. The interpretation was that the two examinations measured the same construct. Hoste suggested that the examinations might have measured some general factor, or general biology ability, or general intellective ability, or ability in written examination, or general scholastic ability, etc. It might be a similar case for EFA at question and sub-

question level. The ability being measured at question level may be could not have been measured at sub-question level.

Again the results of EFA do not have a trend that could be attributed to the influence of teachers' age, qualification, experience, college they attended and school environment in which they were teaching.

4.4 Teachers' perceptions about peer instruction in test construction

During the study two main questions relating to teachers' perceptions about test construction were to be answered. These were: 'To what extent were teachers aware of the need for raising validity evidence of their tests?' and 'What were teachers' perceptions about possibilities of raising validity evidence of their tests through peer instruction in test construction?' Their perceptions were sought at planning for peer instruction workshop, at the end of the workshop and at the end of the 2007 academic year, which was the end of data collection phase of the study. Teachers' perceptions about test construction are given in the sub-sections that follow.

4.4.1 Planning for peer instruction

A questionnaire was administered to the selected sample of teachers to determine the content for peer instruction workshop in test construction. The teachers were requested to show the extent to which they wanted that a topic be included in the content of the workshop.

Table 4.17 Rating of workshop content

		Respondents N = 17 except for item 10, N = 16			Respondents (Percentage)		
	Topic	Rating1	Rating2	Rating3	Rating1	Rating2	Rating3
1	Definition of a test	8	7	2	47	41	12
2.	Description of a test	3	8	6	18	47	35
3.	Purpose of a classroom test	2	3	12	12	18	70
4.	Coverage of classroom test	1	4	12	6	24	70
5.	Writing good test items	0	2	15	0	12	88
6.	Determining order of items	2	5	10	12	29	59
7.	Quality of classroom test	0	2	15	0	12	88
8.	Assembling items for a test	1	6	10	6	35	59
9.	Preparing a marking scheme	1	6	10	6	35	59
10.	Assessing how test items perform **(N = 16)**	1	4	11	6	25	69
	Mean	1.9	4.7	10.3	11.3	27.8	60.9

Rating given in Table 4.17 shows the extent to which the participant wanted the topic to be included in the content for peer instruction, where 1 shows least wanted and 3 shows most wanted. Based on information in Table 4.17 the teachers strongly recommended topics 3 to 10 for content of the workshop. Generally 60.9% of the teachers were in favour of the test construction workshop.

4.4.2 Achievement of objectives in peer instruction workshop

Teachers evaluated peer instruction to find out whether or not its objectives were achieved. They answered the question, 'To what extent were the following objectives achieved by the end of the peer instruction workshop?' The rating of 1 represented least

Table 4.18 Rating of achievement of workshop objectives

		Respondents N = 17			Respondents (Percentage)		
	Objective about	Rating 1	Rating 2	Rating 3	Rating 1	Rating 2	Rating 3
1	Definition of a test	0	0	17	0	0	100
2.	Description of a test	0	6	11	0	35	65
3.	Purpose of a classroom test	1	1	15	6	6	88
5.	Writing good test items	0	3	14	0	18	82
6.	Determining order of items	0	6	11	0	35	65
7.	Quality of classroom test	0	3	14	0	18	82
8.	Assembling items for a test	0	3	14	0	18	82
9.	Preparing a marking scheme	0	4	13	0	24	76
10.	Assessing how test items perform	0	7	10	0	41	59
11.	Applying workshop test construction principles and procedures	0	4	13	0	24	76
	Mean	0.1	3.7	13.2	0.6	21.9	77.5

achieved while 3 represented most achieved. Table 4.18 shows the results of the teachers' rating of achievement of objectives. The opinion of 77.5% of the teachers, according to information in Table 4.18 was that objectives of peer instruction workshop were achieved.

4.4.3 Usefulness of peer instruction workshop

Teachers evaluated the usefulness of the peer instruction workshop and its topics. The question they were answering was, 'To what extent was the peer instruction workshop useful?' The rating of 1 represented least useful and 3 represented most useful. Table 4.19 shows the results of teachers' evaluation of the workshop. Teachers' responses in Table

4.19 suggested that what teachers covered during peer instruction was generally useful. This is an opinion of 81.09% of the teachers.

Table 4.19 Rating of workshop usefulness

		Respondents N = 17			Respondents (Percentage)		
	Topic	Rating 1	Rating 2	Rating 3	Rating 1	Rating 2	Rating 3
1	Definition of a test	3	4	10	18	23	59
2.	Description of a test	2	8	7	12	47	41
3.	Purpose of a classroom test	0	1	16	0	6	94
4.	Coverage of classroom test	0	2	15	0	12	88
5.	Writing good test items	0	2	15	0	12	88
6.	Determining order of items	0	1	16	0	6	94
7.	Quality of classroom test	0	0	17	0	0	100
8.	Assembling items for test	0	3	14	0	18	82
9.	Preparing a marking scheme	0	5	12	0	29	81
10.	Assessing how test items perform	0	6	11	0	35	65
11.	Peer instruction workshop as a whole	0	0	17	0	0	100
	Mean	0.45	2.91	13.64	2.73	17.09	81.09

4.4.4 Relevance of peer instruction workshop

Relevance of topics and peer instruction workshop was another item the teachers evaluated at the end of the workshop. The question they answered was, 'To what extent was the peer instruction workshop relevant?' The rating of 1 represented least relevant and 3 represented most relevant. Table 4.20 shows the results of teachers' evaluation of the relevance of the workshop. Information in Table 4.20 shows that 82.2% of the teachers considered the workshop to have been relevant. In some ways this also might have meant that their expectations as regards their needs in test construction were met.

Table 4.20 Rating of workshop relevance

		Respondents N = 17			Respondents (Percentage)		
	Topic	Rating 1	Rating 2	Rating 3	Rating 1	Rating 2	Rating 3
1	Definition of a test	1	5	11	6	29	65
2.	Description of a test	0	8	9	0	47	53
3.	Purpose of a classroom test	0	3	14	0	18	82
4.	Coverage of classroom test	0	1	16	0	6	94
5.	Writing good test items	0	2	15	0	12	88
6.	Determining order of items	0	1	16	0	6	94
7.	Quality of classroom test	0	1	16	0	6	94
8.	Assembling items for a test	0	4	13	0	24	76
9.	Preparing a marking scheme	0	4	13	0	24	76
11.	Peer instruction as a whole	0	0	17	0	0	100
	Mean	0.1	2.9	14	0.6	17.2	82.2

4.4.5 Degree to which test construction was understood

Teachers also evaluated the degree to which they understood test construction after attending the peer instruction workshop. The question they answered was, ‘To what extent did you understand test construction?’ A rating of 1 represented ‘same as before’, meaning nothing changed, 2 represented ‘slightly better’, 3 represented ‘better’ and 4 represented ‘much better.’ The results of the teachers’ evaluation appear in Table 4.21, which suggested that 65.64% of them learnt more about test construction during the peer instruction workshop.

Table 4.21 Rating of understanding of workshop content

	Respondents N = 17				Respondents (Percentages)			
Topic	Rate 1	Rate 2	Rate 3	Rate 4	Rate 1	Rate 2	Rate 3	Rate 4
Definition of a test	2	2	2	11	12	12	12	64
Description of a test	0	4	6	7	0	24	35	41
Purpose of a classroom test	0	0	5	12	0	0	29	71
Coverage of classroom test	0	0	5	12	0	0	29	71
Writing good test items	0	0	4	13	0	0	24	76
Determining order of items	0	1	4	12	0	6	24	70
Quality of classroom test	0	0	7	10	0	0	41	59
Assembling items for a test	0	0	5	12	0	0	29	71
Preparing a marking scheme	0	2	4	11	0	12	24	64
Assessing how test items perform	0	1	6	10	0	6	35	59
Test construction as a whole	0	0	4	13	0	0	24	76
Mean	0.18	0.91	4.73	11.18	1.09	5.45	27.82	65.64

Key: Rate - rating

4.4.6 Teachers' perceptions about application of test construction skills

As part of the study in-depth interviews were held with individual teachers to establish their perceptions about test construction practice after attending the peer instruction workshop for test construction. Their perceptions were sought on the usefulness and helpfulness of peer instruction, reasons past examination items are used, challenges in test construction after peer instruction and recommendations. Teachers' description of their experiences were interpreted and summarised as given in Table 4.22. See Appendix 4.21 for more details.

Based on information in Table 4.22, teachers generally considered the exercise useful for improving quality of tests and instruction. They were also of the opinion that learners

benefited from the exercise in terms of preparation for the examinations which they wrote as reflected by better performance in MSCE Physical science in 2007 than in 2006.

Table 4.22 Teachers' perceptions about peer instruction

Item	Perception
Usefulness	• Improved quality of tests • Improved instruction
Helpfulness	• High achievement • High teacher confidence in own items
Reason for using past examination items	• Time not available to construct own items • Laziness • Not capable of writing own items • To compare learner performance level on other examinations
Challenges experienced	• Teacher overload • Insufficient teaching resources
Recommendations	• Teacher motivation

To support their perception that the exercise was useful and helpful, information in Table 4.23 shows teachers' claim that in 2007 MSCE Physical Science examination pass rate increased in 12 out of 16 schools representing 75% of the sample schools. Mean pass rate for the schools in 2006 MSCE Physical Science Examination was 53.73% while in 2007 was 64.11%. However, the national pass rate was checked with MANEB. It was found to be higher in 2007 than in 2006. Therefore a higher pass rate for a sample school in 2007 MSCE Physical Science examination than in 2006 might not necessarily be due to the impact of test construction activities of this study alone, if at all it contributed. Most likely, other variables might have played a part.

Table 4.23 Schools' MSCE Physical Science pass-rates

School	2006	2007
1	55.08	77.78
2	94.12	92.31
3	55.56	55.17
4	54.69	59.42
5	77.42	82.26
7	52.56	74.07
8	40.43	66.02
9	56.86	70.06
10	61.54	58.11
11	46.73	69.30
12	46.00	41.53
13	48.75	65.82
14	36.78	47.19
15	37.25	40.00
16	60.53	75.90
17	35.43	50.86
Mean pass rate	**53.73**	**64.11**
National pass rate	**50.67**	**53.99**

Source: The Malawi National Examinations Board

Information given in Table 4.22 also shows that the exercise gave teachers confidence in their items and therefore in their tests as well. This was expressed in terms of reduced use of past examination items. When pre-tests and post-tests were anlysed for past examination items from 1996 to 2006, it was found that the number of past examination items indeed had dropped for the post-tests, as shown in Table 2.24, from 18.63% for T1 to 7.25% for T2 and 21.71% for M1 to 14.57% for M2. The result is in agreement with their opinion but statistically the reduction in the number of past examination items in the post-tests was not significant since the difference between the pairs of the means was not significant (T1 and T2: paired sample t-test, $p > 0.05$; M1 and M2: paired sample t-test, $p > 0.05$). See Appendix 4.14. P-p plot showed that the distribution was normal and therefore

paired sample t-test could be applied. It means that past examination items were equally used in the pre-tests and post-tests.

Table 4.24 Percentage of copied past examination items

	Copied items from past MSCE Physical Science examinations (1996 -2006) (Percentage)			
Teacher	T1	T2	M1	M2
1	90	0	26	58
2	0	0	-	-
3	0	2	66	18
4	8	10	38	3
5	7	0	12	0
7	0	1	*	*
8	0	2	0	0
9	10	9	-	-
10	6	0	5	6
11	55	3	51	29
12	0	4	2	20
13	29	36	22	2
14	19	3	0	6
15	20	0	18	38
16	0	5	37	0
17	54	41	21	24
18	-	-	6	0
Mean	18.63	7.25	21.71	14.57

Key: T1 - end of term 1 tests (pre-tests), T2 - tests from the same domain as end of term 1tests (pre-tests),M1 – past mock tests (pre-tests), M2 - 2007 mock tests (post-tests), (-) - did not present one of thetests, (*) - one of the tests had no marks against items

Some of the items appeared in both the teachers' tests and past MSCE Physical Science examinations as a matter of chance, for example items like 'Define magnification' or 'State Ohm's law'. They can be common items to both teachers' tests and public examinations. This is in a situation where both teachers and examiners for public examinations might be testing the same objectives of the Physical Science syllabus.

Teachers also said that the reason for copying past examination items was lack of time and laziness as given in Table 4.22. In the same Table, it is cited that copying of past

examination items is due to lack of test construction skills. It might be common for teachers who did not go through a teacher training institution as Chavula (2008) says and 2006 Education statistics show. This could also be attributed to teachers who were not properly trained in college (Kadzamira et al., 2004). As given in Table 4.22, use of past examination items was to compare learners' performance in similar examinations.

Table 4.22 also lists teacher overloading as their main challenge. It might be the reason they lacked time to write items of their own. This observation led to obtaining a sample profile of the teachers in order to assess their teaching load and other information. See Appendix 3.1 which shows the profile of the teachers who participated in the study.

Information in Appendix 3.1 shows that teachers were engaged in many activities besides their profession within school and outside. The normal load for a teacher was said to be between 15 and 18 periods a week according to a Head of one secondary school as well as a Head of Physical Science Department of a different secondary school both within Zomba. Going by this information, of the 16 teachers whose details were accessed, 15 teachers were overloaded with a teaching load ranging from 21 to 42 periods a week.

Some of the teachers taught Physical Science from Form 1 to Form 4. It means a teacher taught both JCE and MSCE examination classes. Examination classes are given more attention in schools than non-examination classes. In fact, Physical Science is a practical subject. In spite of claims of lack of resources, it can be very demanding on a Physical Science teacher, in terms of preparation. Generally, Physical Science classes are large. Some of them are double, triple, or even more streams.

In addition to teaching, the teachers carried out administrative responsibilities in their schools, e.g. heading the school, deputising the Head, being a Boarding master or Boarding

mistress, etc. This is normal for secondary school teachers as part of their training. Others had extra lessons with Night schools and private schools. In this kind of arrangement, one would wish to give such work proper attention at the cost of their official load for them to be considered for future engagement by such schools. Teachers justified their engagement with Night or private schools in terms of making extra income besides their obligation to give service.

Table 4.22 again gives lack of resources as another challenge. Teachers argued that large classes needed a lot of resources for teaching. Resources like stationery restrict coverage of the test. A test of good quality might require more stationery. If paper is not adequate the test might be reduced to fewer pages. What is also reduced is the number of questions for the test. The reduction might affect test coverage. What is at stake in this context is content related validity evidence.

As an example, one of the teachers in the sample was writing some of his tests on the chalk board for lack of stationery for a class of 70 pupils. In such a situation a teacher might not be motivated to sample the test domain adequately. Similarly, chemicals and equipment determine whether or not a practical paper is included in a test or furthermore, which practical question to include in the practical paper. Hence, content related validity evidence is at stake. Table 4.25 shows teachers' frequency of administration of a practical test paper during the study.

Based on information shown in Table 4.25 many teachers were not able to administer a Physical Science practical paper with each test they administered during the study. In the absence of a test blueprint for M1 it was difficult to tell whether or not it had a practical component. Perhaps the reasons were what the teachers stated during the interview that

they did not have enough resources for the practical component. They might have been reserving the resources since it meant conducting five practical tests including a 2007 MSCE Physical Science practical paper for national the examination. It could also have been that they were overloaded, lazy and lacked motivation as expressed during the discussions.

Table 4.25 Frequency of Physical Science practical tests

Teacher	T1	T2	M1	M2	Total
1	1	0	0	1	2
2	1	0	-	1	2
3	0	0	0	0	0
4	0	1	0	1	2
5	0	0	1	1	2
7	1	0	0	1	2
8	0	0	0	1	1
9	0	0	-	1	1
10	0	0	0	1	1
11	0	0	0	1	1
12	1	0	0	1	2
13	0	0	0	1	1
14	0	0	1	1	2
15	0	0	0	0	0
16	0	0	0	1	1
17	0	0	0	0	0
18	-	-	0	1	1

Key: (-) – tests not submitted

It was also noted from the interviews, as given in Table 4.22 that teachers would like to be motivated to teach. This was expressed in many different ways. The most popular way suggested was recruitment of teachers to reduce their overload. Other suggestions included frequent in-service training, better salaries and provision of instructional materials. Lack of motivation could be a major reason for teachers not applying their test construction skills. As long as there are no incentives in teaching, teachers will continue to be overloaded as a way of bridging the gap of incentives.

Much as teachers developed confidence in their items the problem of copying past examination items will persist in schools considering that they continue to have no time for writing items of their own. They would also wish to compare performance of learners on test items from other examiners. Laziness and lack of motivation as long as they exist in schools copying of past examination items will persist.

4.4.7 Summary: Teachers' perceptions about peer instruction

Based on responses to questionnaires and discussions with them, teachers were in favour of the peer instruction workshop in test construction. Their responses also indicated that the workshop objectives had been achieved, workshop content was useful and that they learnt more about test construction from the workshop. Teachers also had difficulties with test construction, which were expressed in terms of challenges and recommendations.

Several interpretations are made about teachers' perceptions of peer instruction. As regards to training needs assessment on a planning peer instruction workshop, teachers were aware of their CPD needs in test construction. Responding to achievement of objectives, teachers showed a sense of appreciation and satisfaction that their expectations were met. This appreciation and satisfaction included definition and description of a test, topics 1 and 2, which had been rated lowly for inclusion as part of workshop content. It might have been an indicator of adjustment for proper conceptualisation of what a test is and its description.

Regarding usefulness, relevance and understanding of the topics, it is interpreted that the teachers found the CPD session in test construction vital for improving quality of their class tests and instruction. This is not surprising because the content of the workshop in test

construction was tailored to their needs. These interpretations were supported by teachers' further reactions like, asking for another meeting, where their progress would be reviewed and participants sharing knowledge and skills acquired from peer instruction workshop in test construction with other teachers in school.

What was learnt from the discussions with teachers at the end of the data collection phase was that teachers benefited from peer instruction. The value they attached to the peer instruction workshop and its content was similar to that one they attached to its application during teaching as well as planning and delivery. However, challenges they experienced and recommendations put forward for improvement indicated that the teachers had difficulties with effective application of test construction skills acquired from the peer instruction workshop. Possibly the source of difficulties relate to overloading, lack of resources and low motivation.

4.5 Conclusion

The study has found that, according to item analysis results, the percentage of good and excellent items combined, with respect to item discrimination, is generally high in teachers' pre-tests and post-tests. This was interpreted to mean that teachers were able to identify good and excellent items for their pre-tests and post-tests to a greater extent. Therefore, teachers had knowledge and skills for supplying good and excellent items for their tests even before they attended peer instruction for test construction. Similarly, reliability was also generally high for their pre-tests and post-tests. Again, teachers were able to construct good tests in terms of reliability even before attending peer instruction.

In content related validity evidence, the study has found that the proportion of teachers' items which were item relevant or had construct relevance in the pre-tests and post-tests

were also equally high too. It reflects teachers' capability of supplying items of their tests from a defined test domain. The study also shows that item cognitive representativeness that the percentage of recall and comprehension items were each higher than the proportion of higher order items in both pre-tests and post-tests. Teachers' practice of supplying more recall and comprehension items than higher order items for their tests did not improve with peer instruction. It might mean that use of more recall and comprehension items in a test is a deep rooted practice. The proportion of representative items was found to be more for pre-tests of T2 form. It must be the effect of peer instruction on item representativeness. Teachers learnt, through peer instruction, how to construct tests which are more item representative.

It has also been established, through this study, that construct related validity evidence increased for both T2 and M2 at question level also as an effect of peer instruction. However, there was negligible increase in construct validity evidence for T2 and M2 at sub-question level. The explanation for the observed disparity most probably lies in the characteristics of factor analysis. It might be behaving differently when the same data is rearranged, in addition to increased inter-correlation of sub-questions.

Based on this study again, it is found that teachers were aware of their professional gap in test construction, in addition to the need to bridge the gap. The teachers considered the short term workshop to have been beneficial in building their capacity in test construction. However, teachers' narratives during discussions showed that there were some challenges to effective application of what they learnt during the workshop. The challenges were narrated as overloading, lack of motivation and resources.

What is learnt from this study is that there is potential for raising validity evidence of Physical Science tests through teachers' peer instruction. However, this is with respect to item representativeness in content related validity evidence and also in construct related validity evidence. Comparative results of both content and construct related validity evidence between pre-test and post-test pairs are consistent with teachers' perceptions about effectiveness of peer instruction on test construction.

Effectiveness of peer instruction as a methodology for learning as found in the study is also consistent with what Lasry (2006), Cahyadi (2003), Fagen, et al. (2002), Crouch & Mazur (2000) and Hake (1998) found in Physics courses and Cortright, et al. (2005) found in a Physiology course. Therefore peer instruction would facilitate learning not in academic courses only but in a CPD class as well.

As an innovative learning strategy peer instruction applies learning theories of John Dewey, Jean Piaget, Levi Vigotsky, Jerome Bruner and Malcom Knowles. Individuals learn more if they are given an opportunity to be actively engaged in building own knowledge on the experience they already have and through sharing with other learners (Epstein & Ryan, 2002; Emand & Fraser, 2002; Riddle & Dabbagh, 1999). Since the teachers engaged in peer instruction were adults, what was instrumental to their learning were their characteristics of 'interest in learning what is of immediate relevance to their jobs' and also 'problem centred learning' as well (Cranton, 1989). The findings of this study support these theories.

Peer instruction has had a positive impact on learning in a CPD class, in the case of this study. This might be an indicator of a possibility of a similar impact in other CPD programmes that use the same instructional strategy, conducted by the Ministry of

Education Science and Technology in Primary Schools in Malawi. This study serves as an eye opener to CPD programmes in Primary Schools and at any other level of the education system, that peer instruction can enhance meaningful learning of teachers in such programmes. There is need though for another study whose findings could be more generalizable to all forms of CPD programmes in the education system in Malawi.

CHAPTER 5:

CONCLUSIONS, IMPLICATIONS AND RECOMMENDATIONS

5.0 Introduction

Chapter 5 gives a synopsis of the study with reference to the objectives, research questions, findings and conclusions. The implications are discussed together with recommendations for further research. The recommendations are made for all who play a role in effective delivery of instruction and future researchers.

5.1 Conclusions

The main purpose of this study was to explore the feasibility of raising validity evidence of Physical Science tests through teachers' peer instruction in test construction. The questions which the study addressed were:

a. Were teachers' post-test items an equally relevant and representative sample of the test domain as pre-test items?
b. Did teachers' post-test items equally measure learners' cognitive ability levels as pre-test items?
c. Were the means of percentage total variances explained by common factors between the teachers' post-tests and pre-tests the same?

f. To what extent were teachers aware of the need for raising validity evidence of their tests?

g. What were teachers' perceptions about possibilities of raising validity evidence of their tests through peer instruction in test construction?

The study employed a mixed methods approach. Qualitative methods complemented quantitative methods which applied a pre-test and post-test one group experimental design. The rationale was to gain an in-depth understanding of the possibility of increasing validity evidence of teacher made tests through peer instruction, and underlying contextual issues.

Teachers were oriented in MSCE Physical Science test construction skills through a peer instruction workshop. They were expected to apply these skills during instruction. Teachers' pre-tests and post-tests were evaluated to determine whether or not quality of their tests had improved after attending the peer instruction workshop in test construction. An increase in content and construct related validity evidence were to be the indicators of this improvement. The study also assessed teachers' perceptions about test construction through questionnaires and in-depth interviews. This was done in order to determine their test construction experience. Teachers' experiences with test construction were crucial for in-depth understanding of the phenomenon under investigation. To this effect, a number of conclusions have been reached in this study and are discussed in this section.

The significance of this study was that if it were possible to raise validity evidence of teacher made tests through peer instruction then peer instruction could provide a cost-effective strategy for improving classroom tests, and in turn instruction and achievement in schools.

5. 1.1 Item relevance and representativeness

Item relevance for teachers' pre-tests and post-tests was equally high in both sets of tests. It means teachers were capable of sourcing relevant items for a test within a given test domain even before peer instruction. This would be expected since a teacher's items would be testing specific objectives of topics outlined in a syllabus. Chances of irrelevant items would be minimal if a teacher refers to topic objectives of the test domain when constructing their tests. The finding might also mean that teachers closely use the syllabus when constructing their tests.

In terms of item representativeness, the tendency was that items for T2 were more representative than items of T1. It was therefore possible to increase this aspect of content related validity evidence of teachers' Physical Science tests through peer instruction. The interpretation of this was that teachers' deficiencies were in constructing tests which adequately sample the test domain of interest. Therefore, content validity of their tests was low because of item representativeness.

5.1.2 Cognitive level of items

Teachers' pre-test and post-test items equally tested recall and comprehension levels more than higher order levels generally. This is the argument which Mwanza and Kazima (2000) as well as Bregman and Bryner (2003) have about classroom tests. Items are not equally distributed across cognitive levels. Teachers' test items are generally of low order. This seems to be a deep rooted practice for teachers since the situation did not improve after the teachers attended the peer instruction workshop. The degree to which items test cognitive levels is an issue of item representativeness in a test, with respect to the level of

abilities being tested. Therefore it lowers content related validity evidence of a test in that items which demand higher cognitive abilities are not well represented in the test.

The tendency for teachers to test lower cognitive levels more than higher levels suggests many things. The experience is that, higher order test items are difficult to construct and they are time demanding on construction.

5.1.3 Proportion of variance due to common factors

From the results, construct related validity evidence of teachers' tests increased after the peer instruction workshop in test construction. This is in agreement with item representativeness as observed for T2. The conclusion therefore was that construct related validity evidence of teacher made tests could be increased through peer instruction, so long as construct validation is done at question level. More needs to be done to establish what happens at sub-question and item levels regarding construct related validity evidence.

5.1.4 Planning and delivery of peer instruction workshop

Two main conclusions could be drawn about teachers' perceptions of test construction based on the results of evaluation done at planning and delivery of the peer instruction workshop. Firstly, teachers were aware of their professional gap in test construction and the need for redressing it, which was reiterated during discussions with them. When asked what recommendations they had for improving quality of test construction in schools, the response of 69% of the teachers was 'motivate the Physical Science teacher with regular in-service training'. This is in a way asking for more CPD activities to close the existing gap of skills for test construction.

Secondly, teachers considered it possible to improve test construction knowledge and skills, and in the long run, improvement of their tests through peer instruction in test construction. It is important to note that the teachers were involved in planning for peer instruction. Peer instruction might have been highly rated on the grounds that as adult learners, their involvement in planning its content made peer instruction more relevant to their needs (Cranton, 1989). Consequently, they became motivated intrinsically to acquire more test construction skills from their peers (Palmer, 1978 in Rubin, 1978). Coupled with the principle that an individual must construct their own knowledge (Redish, 1994), the adult learner being self-directed might have found it to be an opportunity to interact with other teachers in order to learn more from them on difficulties they have with test construction.

5.1.5 Teachers' perceptions about application of test construction skills

Teachers' perceptions about peer instruction in test construction as drawn from the discussions with the teachers were consistent with their perceptions captured at the end of the workshop in test construction. Teachers considered peer instruction useful and helpful for improving quality of their class tests. They believed that they had benefited from it in terms of improved quality of tests while learners had their achievement improved. The teachers' perceptions in this regard suggested that class tests could be improved with peer instruction in test construction, as an intervention.

Teachers' perceptions also showed that their class tests could have improved more if it were not for challenges they experienced with test construction, i.e. overloading, lack of resources and low motivation. The challenges teachers experienced could also be

understood better from their recommendations. Certainly, this can be a way forward for improving classroom tests in schools.

5.1.6 Summary: Conclusions

Based on the results of this study it was possible to improve content related validity evidence of the teachers' tests through peer instruction as observed from T2. Item representativeness contributed to low content related evidence in T1. As a result it increased for T2 with peer instruction. Teachers' tendency to test lower cognitive levels more than higher cognitive levels prevailed in post-tests. It was an indicator that this tendency could not be improved through peer instruction.

Construct related validity evidence was possible to increase through peer instruction at question level. More investigations are required to find out what happens to construct related validity evidence at sub-question and item levels.

In terms of teachers' perceptions, it was noted that they were keen to redress their gap in test construction. After going through the peer instruction workshop in test construction, their perceptions indicated satisfaction with delivery of instruction. However, it was revealed that there were challenges in effective application of what they learnt during peer instruction in test construction.

The conclusions were consistent with the findings about peer instruction as discussed in Chapter 2 that it was effective for learning Physics concepts (Lasry, 2006; Cahyadi, 2003; Fagen, et al., 2002; Crouch & Mazur, 2000; Hake, 1998). It was also effective in enhancing meaningful learning in a physiology class (Cortright, et al., 2005). Having found that it was possible to improve teachers' capacity in constructing tests of high

validity evidence through peer instruction, it is an indicator that peer instruction has the potential for meaningful learning in CPDs as well.

5.2 Implications of the findings

Conclusions made on the results have many implications on the education system. The sections that follow discuss these implications.

5.2.1 Validity evidence

It was evident in this study that low content related validity evidence of teachers' tests was a contribution of item representativeness more than item relevance. This included cognitive depth the items tested since most items were of lower cognitive levels. In this regard, improving teachers' capacity in constructing tests of high content related validity evidence, emphasis should be on assisting the teachers to construct tests which adequately sample the domain of interest besides a fair representation of the cognitive depth which items should test. CPDs should focus on how to adequately sample both a test and cognitive domain.

Consistent increase of content and construct related validity evidence was noted for T2. It might have been an indicator that increasing content related validity evidence in a test, contributes to an increase in construct related validity evidence. This is a plausible observation since test content defines underlying constructs which would account for an examinee's performance in that test.

Results also showed that teachers' construct related validity evidence could be increased with peer instruction. However, procedures should be investigated for directly

improving construct related validity evidence of classroom tests. EFA procedures might be demanding on a teacher. They involve pre-testing of the items. Perhaps this is where attention should be given to improving teachers' capacity in constructing tests which adequately sample a test domain.

5.2.2 Teachers' perceptions of test construction

It was noted that teachers were aware of their professional gap in test construction and need for redressing the gap. Based on education statistics of the Ministry of Education and Vocational Training (2006), many more teachers in the education system should have training needs in test construction. Capacity building for them as well is necessary. Since teachers considered content and peer instruction workshop in test construction to have been most relevant and most useful, similar content and delivery of CPD in test construction might be relevant, useful and cost-effective for building capacity in test construction of other teachers for the improvement of quality of their tests and instruction. This could also assist to establish further the potential of peer instruction as an effective mode of delivery for CPDs.

Teachers' perceptions relating to application of test construction knowledge and skills acquired from the peer instruction workshop reflected that the teachers experienced challenges in applying test construction knowledge. Their recommendations were perceived to be measures intended to reduce difficulties they experienced with test construction during teaching. Other teachers going through a similar CPD would likely have the same or similar challenges. It is necessary, therefore, that in order to improve quality of classroom tests, consideration should also be given to reducing these challenges

which include overloading, lack of teaching material and lack of incentives. These lead to teachers' low motivation and morale.

Bennell and Akyeampong (2007) in their study identified these challenges as causes of teachers' low job satisfaction in Sub-Saharan Africa and South Asia. Therefore a discussion about teachers' challenges and recommendations during discussions with teachers in this study was a discussion of their motivation and morale to apply the acquired test construction knowledge and skills during teaching.

Kadzamira (2006), in a case study of Malawi on teacher motivation which includes incentives says that low morale has reduced teacher performance. Consequently, teachers find excuses to absent themselves from school for secondary activities to supplement their income. Most teachers participating in this study engaged in private tutoring, which is a symptom of poor teacher motivation and low morale for teaching. See Appendix 3.1. It might have compromised their commitment to official duties including application of test construction knowledge and skills acquired during the peer instruction workshop. The end result of it might have been low teacher output (Adelabu, 2005) as implied by Kadzamira.

Teachers' recommendation for better salaries confirms what most studies on teacher motivation in Africa and Malawi in particular have found that teachers are poorly paid. They receive little or no allowances in their services, promotion is not regularly done and career progression is limited. Doctors, Jambane, Marsh and Ngomane (2009) in their study on education workers' motivation and morale in Mozambique found that as a result of low motivation teachers did not have time to prepare for their classes adequately, which might include using recall and comprehension items in their tests. In a situation like that teachers become resistant to new teaching methodologies and other innovations (Bennell &

Akyeampong, 2007). The question was whether or not it applied to this study as well. If at all it did, then the results of this study should have been better with high teacher motivation.

5.2.3 Summary: Implications

Content and construct validation study done on improving quality of teachers' tests showed that peer instruction has the potential for increasing validity evidence of the tests. This was also reflected in teachers' perceptions about peer instruction and test construction. However, teachers perceptions relating to challenges experienced during this exercise were indicators that CPDs alone may not be sufficient means for improving quality of tests in schools. CPDs should be complemented by measures which would improve teachers' motivation and morale to teach.

5.3 Recommendations for further research

The focus of this study was improvement of quality of classroom tests for improved instruction and achievement in MSCE Physical Science. In order to achieve this, consideration should be given to a number of supporting issues as raised in this study. Attention should be given to recommendations given in this section.

5.3.1 EFA at different question levels

During the study EFA at question level and sub-question level gave contrasting results about common factors and percentage total variance explained by the common factors. Extraction of factors at item level was not successful for most of the tests. An EFA study

with a larger sample size, 500 – 1000 subjects, could be done to compare results at question level, sub-question level and item level. Such a study could clarify what Physical Science constructs are measured at those question levels. However, it might be difficult to get a suitable sample size for such an exercise from a single school. Sometimes scores of candidates for public examinations, with permission from MANEB, could be used to provide examination data for a suitable sample size.

5.3.2 Attributes underlying performance in Physical Science

This study did not establish attributes which underlie performance in MSCE Physical science tests. Analysis of items which cluster together in EFA could be done to identify such attributes. This could be followed by CFA_2 to verify attributes proposed by EFA. A sample size of at least 200 learners or candidates could be used in such an investigation, if it is a public examination.

5.3.3 Evaluation of item representativeness of pre-tests and post-tests

During this study, item representativeness was compared between T1 and T2. Another study could be done to compare item representativeness of pre-test and post-test pairs of tests which have practical components. It could apply the same procedures which were applied for T1 and T2 in this study.

5.3.4 Cognitive depth of teachers' test items

It was not possible to improve distribution of items across cognitive levels of teachers' tests with peer instruction. It would be necessary to have empirical information regarding

teachers' tendency of constructing Physical Science tests whose items generally lack cognitive depth. The sample should include a mixture of teachers with high and low experience, and both trained and untrained.

5.3.5 Incentives as intervention

Teachers' recommendations reveal constraints in as far as improvement of teacher made tests in schools are concerned. Their recommendations were for various forms of motivation. Using a sample which is assumed to have test construction skills, investigations could be made to find out whether or not teacher made tests could be improved further by removing some of the factors which make teachers less motivated to teach. These factors include overloading, promotion, salaries and allowances. For example, the question to be answered in this context example could be, 'Does providing a specific incentive to teachers increase validity evidence of their tests?' The type of incentive should be mentioned instead of being general when it comes to real proposals.

REFERENCES

Abdi, H. (2003). Factor rotations in factor analysis. Accessed on 2 March 2010 from http://www.utdallas.edu/~herve/Abdi-**rotation**s-pretty.pdf

Adelabu, M. A. (2005). Teacher motivation and incentives in Nigeria. Unpublished research report. Accessed on 17 October 2007 from http://www.pdfgeni.com

Addock, R. & Collier, C. (2001). Measurement validity: A shared standard for qualitative and quantitative research. *The American Political Science Review, 95(3), 529 - 546.*

AERA, APA & NCME. (1999). Standards for educational and psychological testing. Washington, DC: American Educational Research Association.

Anastasi, A. (1986). Evolving concepts of test validation. *Annual Review of Psychology*, *37*, *1-15.*

Angoff, W. H. (1988). Validity: An evolving concept. In H. Wainer and H. I. Braun (Eds). Test Validity (19 – 33). New Jersey: Lawrence Erlbanm Associates.

APA.(2002). Ethical principles of psychologists and code of conduct. Accessed on 7 June 2008 from http://www.apa.org/

Australian Government.(2008). Human Research Ethics Handbook. National Health and Medical Research Council. Accessed on 31 May 2008from http://www.nhrmc.gov.au/

Atherton, J. S. (2005). Learning and teaching: Knowles' andragogy: an angle on adult learning. Accessed on 21 July 2008 from http://www.learningandteaching.info/learning/knowlesa.htm

Bennell, P. & Akyeampong, K. (2007). Teacher motivation in Sub-Saharan Africa and South Asia. Unpublished DFID Educational papers. Accessed on 18 October2007 from http://www.dfid.gov.uk/

BERA. (2004). Revised ethical guidelines for educational research. Accessed on 7 June 2008 from http://www.bera.ac.uk

Black, P. & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in education, 5(1), 7 – 76.*

Boston, C. (2002). The Concept of formative assessment. Practical Assessment, Research and Evaluation, 8(9), 1 - 6. Accessed on 2 February 2010 from htt://PAREonline.net/getvn.asp?v=8&n=9

Breg, B. L. (1998). Qualitative research methods for the social sciences (3rd edition.). USA. Allyn and Bacon.

Bregman, J & Bryner, K. (2003). Quality of secondary education in Africa. Association for the Development of Education in Africa. Unpublished. Accessed on 2 January 2006 from http://www.adeanet.org

Bude, U. & Lewin, K. (Eds.). (1997). Improving test design. Vol.1 – Constructing test instruments, analyzing results and improving assessment quality in Primary schools in Africa, 6 - 11. Bonn: German Foundation for International Development Education, Science and Documentation Centre (ZED).

Cahyadi, V. (2003). The effect of interactive engagement teaching method to student understanding of introductory Physics at the Faculty of Engineering, University of Surabaya, Indonesia. *Higher Education Research and Development Journal, 23(4), 455 – 464.*

Chakwera, E.W. J. (2004). Content validity of independently constructed curriculum based examinations. Unpublished doctoral dissertation, School of Education University of Massachusetts, Amherst.

Chakwera, E. W. J. (2005). Improving examination performance through teacher capacity building in assessment. In Educational assessment in Democracy: Challenges, opportunities and prospects. Unpublished paper presented at the 3rd Sub-regional Conference on assessment, Zomba, Malawi. Board, Behind Old Parliament Building, P.O. Box 191 Zomba Malawi.

Chavula J. (2008, September 3). Ministry to regulate teaching profession. The Daily Times, p. 3.

Clark, C. A. (1959). Developments and applications in the area of construct validity. *Review of Educational Research, 29 (1), 84 – 104.* Educational and Psychological Testing. (1959, February), 84 – 105. Accessed on 2 July 2006 fromhttp://www.jstor.org/

Conner, M. L. (2005). "Andragogy and pedagogy." Ageless Learner, 1997-2004. Accessed on 7 December 2006 from http://agelesslearner.com/intros/andragogy.html

Conway, J. M. & Huffcut, A. I. (2003). A review and evaluation of exploratory factor analysis practices in organizational research. *Organizational Research Methods, 6 (2), 147 – 168.*

Corbin, J. (1990). Basics of qualitative research. California. Sage Publication.

Cortright, R. N., Collins, H. L. &DiCarlo, S. E. (2005, March). Peer instruction enhanced meaningful learning: ability to solve novel problems. *Advances in Physiology Education, 29, 107 – 111.*

Costello, A. B. & Osborne, J. W. (2005). Best practice to exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical Assessment, Research and Evaluation,10 (7).* Accessed on 3 June 2006 from http://pareonline.net/getvn.asp?v=10&n=7

Coughlin, M. A., & Knight, W. (2007). Exploratory factor analysis. Accessed on 22 March 2007 from http://www.airweb.org

Cranton, P. (1989). Planning instruction for adult learners. Toronto: Wall and Emerson.

Creswell, J. W. (1998). Qualitative inquiry and research design: Choosing among five traditions. California: SAGE Publications.

Crocker, L., &Algina, J. (1986). Introduction to classical and modern test theory. Belmont, California: Wardsworth Group.

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.*

Cronbach, L. J. (1971): Test validation. In R. L. Thorndike (Ed), Educational measurement *(*2nd ed.), (pp. 443 – 507). Washington, DC: American Council on Education.

Cronbach, L. J. & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52 (4), 281 – 302.*

Cronbach, L. J. & Shavelson, R. J. (2004). My current thoughts on coefficient alpha and successor procedures. *Educational and Psychological Measurement, 64 (3), 391 – 418.*

Crouch, C. H. & Mazur, E. (2001). Peer instruction: Ten years of experience. *American Journal of Physics, 69(9), 970 – 977.*

Cuseo, J. (2008). Evidence supporting the positive impact of peer tutoring. Accessed on 5 March 2008 from http://www.peerassistance.bigstep.com/

Darlington, R. B. (2007). Factor analysis. Accessed on 25 December 2005 from http://www.psych.cornell.edu/Darlington/factor.html

Darton, R. A. (1980). Rotation in factor analysis. *The Statistician, 29(3)), (167 – 194).*

Denzin, N. K. & Lincoln, Y. S. (1998). Collecting and interpreting collective material. California. Sage Publication.

Department of Health, Education and Welfare, (1979).The Belmont report: Ethical principles and guidelines for the protection of human subjects of research. Accessed on 31 May 2008 from http://www.hhs.gov/ohrp/humansubjects/guidance/belmont.html

Department of Health and Human Service (2005). Regulations and ethical guidelines. Accessed on 24 May 2008 from http://ohrs.od.nih.gov/

Doctors, S., Jambane, S., Marsh, R. & Ngomane, J. (2009). Listening to teachers: The motivation and morale of education workers in Mozambique. Accessed on 17 October 2007 from http://www.vso.org.uk

Downing, S. M. (2003). Validity: on the meaningful interpretation of assessment data. Medical Education, 37, pp. 830 – 837. Accessed on 25 October 2006 from http://www.ncbi.nlm.nih.gov/

Emand, N. I. & Fraser, S. (2006). Educational theory of John Dewey (1859 – 1952). Accessed on 8 December 2006 from http://www.newfoundations.com/GALLERY/Dewey.html

Education Broadcasting Corporation. (2004). Workshop: Constructivism as a paradigm for teaching and learning. Accessed on 18 July 2008 from http://www.thirteen.org

Ender, P. (1998). Multivariate analysis: Exploratory factor analysis. Accessed on 29 December 2009 from http://www.philender.com/courses/multivariate/notes2/fa.html

Epstein, M. & Ryan, T.(2002). Constructivism. Accessed on 18 July 2006 fromhttp://tiger.towson.edu/users/

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C. &Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological methods, 4(3), 272 – 299.*

Fagen, A. P. Crouch, C. H. & Mazur, E. (2002, April). Peer instruction: Results from a range of classrooms. *The Physics teacher, 40, 206 – 209.*

Fernandez, G. (2003). Data mining using SAS applications. Accessed on 2 March 2010 from http://books.google.com/books?id=NfmGsy_X44IC&pg=RA1-

Field, A. (2005). Factor analysis using SPSS. Accessed on 12 October 2005 from http://www.stasticshell.com

Froman, R. D. (2001). Elements to consider in planning the use of factor analysis. *Southern Online Journal of Nursing Research, 2(5), 1 - 2.*

Garson, D.G. (2006). Factor analysis. Accessed on 17 August 2006 from http://www2.chass.ncsu.edu/garson/pa765/factor.html

Gronlund, N.E. (1988). How to construct achievement tests (4th ed.). New Jersey: Prentice-Hall.

Guilford, J. P. (1946). New standards for test evaluation. *Educational and Psychological Measurement, 6, 427 – 439.*

Guion, R. M. (1977). Content validity – The source of my discontent. *Applied Psychological Measurement, 1(1), 1 – 10.*

Guion, R. M. (1978). Scoring of content domain samples: The problem of fairness. *Journal of Applied Psychology, 63(4), 499 – 506.*

Guion, R. M. (1980). On Trinitarian doctrines of validity. *Professional psychology, 11(3), 385 – 398.*

Hake, R. R. (1998). Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses. *American Journal of Physics, 66(1), 64 - 74.*

Heller, P., Keith, R & Anderson, S. (1992). Teaching problem solving through cooperative grouping Part 1: Group versus individual problem solving. *American Journal of Physics, 60 (7), 627 - 636.*

Henson, R.K. & Roberts, J. K. (2006). Use of exploratory factor analysis in published research: Common errors and some comment on improved practice. Accessed on 9 December 2008 from http://epm.sagepub.comat

Hillel, J. (2005). Physics education research – A comprehensive study. A Thesis submitted in partial fulfillment of the requirements of the degree of Bachelor of applied science. University of Toronto. Accessed on 11 April 2008 fromwww.upscale.utoronto.ca

Hinkle, D. E., Wiersma, W. &Jurs, S. G. (1998). Applied statistics for the Behavioural science (4thed). Boston: Houghton Mifflin Company.

Hogatry, K. Y., Hines, C. V., Kromrey, J. D., Ferron, J. M. & Mumford, K. R. (2005). The quality of factor solutions in exploratory factor analysis: The influence of sample size, communality and over determination. *Educational and Psychological Measurement, 65(2), 202 – 226.*

Hopkins, K.D. (1998). *Educational and Psychological Measurement and Evaluation*. Boston: Allan and Bacon.

Hoste, R. (1982). The construct validity of some Certificate of Secondary Education Biology Examination: The evidence from factor analysis. *British Educational Research Journal, 8(1), 31 – 43.*

Jackson, S. L. (2009). Research methods and statistics: A critical thinking approach. Accessed on 20 November 2009 from http://books.google.com/books*?*

Johnson, R. (1984). Elementary Statistics (4th Edition). Boston: Duxbury Press

Johnson, D., Hayter, J. & Broadfoot, P. (2000). The quality of learning and teaching in developing countries: Assessing literacy and numeracy in Malawi and Sri Lanka Education Research Paper No.41, 2000, 90. Accessed on 18 October 2007 from http://www.eric.ed.gov/

Kadzamira, E.C., Moleni, C, Kholowa, F., Nkhoma, M., Zoani, A, Chonzi,R. et al. (2004). Student testing and assessment reform: A consultancy report presented to Ministry of Education and Human Resources Development. (Available at CERT, Chancellor College, P.O. Box 280, Zomba, Malawi).

Kadzamira, E. C. (2006). Teacher motivation and incentives in Malawi. Unpublished research report. Accessed on 17 October 2007 from http://.www.pdfgen.com

Kaira, L. (2003). Malawi teachers' knowledge of and attitudes towards Standardized Tests. Unpublished master's thesis, School of Education University of Massachusetts, Amherst.

Kane, M. T. (1992). An argument based approach to validity. *Psychological Bulletin, 112(3), 527 – 535.*

Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement, 38(4), 319 – 342.*

Kane, M. T. (2002, Spring). Validating high stakes testing programs. *Educational Measurement: Issues and Practices, 21(1), 31 – 41.*

Kang, S. J. & Park, J. H. (2004). Validity: A unified concept. Unpublished paper presented to AAHPERD National convention, March 30 – April 3, 2004. Accessed on 7 June 2006 from http://www.mtsu.edu/seind10.pdf

Kellaghan, T. & Greaney, V. (2003). Monitoring performance: Assessment and examinations in Africa. Association for the Development of Education in Africa. Accessed on 29 November 2007 from http://www.adeanet.org

Kerka, S. (2002). Teaching adults: Is it different? Myths and realities no. 21. Accessed on 7 December 2006 from http://www.calpro-online.com/ERIC/

Kline, P. (2009). The handbook of psychological testing. Accessed on 15 January 2010 http://www.amazon.co/

Kushnir, L. P. (2006). Teaching by questioning – Peer instruction: An innovative approach to teaching and learning. Teaching and learning symposium. University of Toronto. Accessed on 11 April 2008 from http://www.utoronto.ca/

Lane, S. & Stone, C. A. (2002). Strategies for examining the consequences of assessment and accountability programs. *Educational Measurement: Issues and practice, 21(1), 23 – 30.*

Lasry, N. (2006). Implementing Peer instruction in Cegep**.** Retrieved on 11 April 2008 from http://www.cdc.qc.ca/parea/

Lebacqz, K. (1980). Beyond respect for persons and beneficence: Justice in research. *IRB- A Review of Human Subject Research, 2(7), .1 - 2.* Accessed on 31 May 2008 from http://www.jstor.org/action/

Lieb, S. (1991). Principles of adult learning. Accessed on 11 December 2006 from http://honolulu.hawaii.edu/intranet/committees/

Lindboe, T. A. (1998). The effectiveness of Peer instruction in the learning environment of low-achieving undergraduate mathematics students. Accessed on 11 April 2008 from http://proquest.umi.com/pqdlink/?=738163921&Fmt=7&clientld

Linn, R. L. (Ed.) (1989). Educational measurement (2nd ed.). New York: American Council on Education and Macmillan.

MacCallum, R. C., Widaman, K. F., Zhang,S. & Hong, S. (1999). Sample size in factor analysis. *Pyschological methods, 4(1), 84 – 89.*

MacCallum, R. C., Widaman, K. F., Preacher, K. J. & Hong, S. (1999). Sample size in factor analysis: The role of model error. *Multivariate Behavioural Research, 36(4), 611- 637.*

Malawi Government. (1999). The Constitution of the Republic of Malawi. Gothenburg-Sweden: NovumGrafiska AB.

Mazur Group. (2008). Peer instruction. Accessed on 7 March 2008 from http://mazur-www.harvard.edu/research/detailspage.php?rowid=8

McDermott, L. C. (1991). Millikan Lecture 1990: What we teach and what is learned – Closing the gap. *American Journal of Physics, 59(4), 301 – 315.*

McDermott, L. C. (1993). Guest comment: How we teach and how students learn – A mismatch? *American Journal of Physics 61(4), 295 – 298.*

Meltzer, D. E., & Manivannan, K. (2002). Transforming the Lecture-hall environment: The fully interactive physics class. *American Journal of Physics 70(6), 639 - 654.*

Messick, S. (1989a). Validity. In Robert L. Linn (Ed.). Educational measurement (2nd ed., 13-103). New York: American Council on Education and Macmillan.

Messick, S. (1989b). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher, 18(2), 5 – 11.*

Messick, S. (1995). Standards of validity and validity of standards in performance assessment. *Educational measurement: Issues and practice, 14(4), 5 – 8.*

Ministry of Education and Vocational Training, (2006). Education statistics, 2006. (Available from Department of Education Planning, Ministry of Education and Vocational Training, Private Bag 328, Capital City, Lilongwe 3, Malawi).

Mosier, C. I., (1947). A critical examination of the concepts of face validity. *Educational and Psychological Measurement, 7, 191- 205.*

Moss, P. A. (1992). Shifting conception of validity in education assessment: Implications for performance assessment. *Review of Educational Research, 62(3), 229 – 258.*

Moss, P. A. (1994). Can there be validity without reliability. *Educational Researcher, 23(2), 5 - 12*

Moss, P. A., (1995).Themes and variations in validity theory. *Educational measurement: Issues and practice, 14(2), 5 – 13.*

Mundfrom, D. J., Shaw, D. G. & Ke, T. L. (2005). Minimum sample size recommendations for conducting factor analysis. *International Journal of testing*, 5(2), 159 – 168.

Mwanza, S. J. & Kazima, M. (2000). Towards examination reform: A comparison of school assessment by the Malawi National Examinations Board (MANEB) in Mathematics and Science. (Available at Malawi National Examinations Board, Behind Old Parliament Building, P.O. Box 191 Zomba, Malawi).

NASW. (1999). Code of ethics of the National Association of Social Workers. Accessed on 7 June 2008 from http://www.socialworkers.org/pubs/code/code.asp

Narman, A. (1995). Manual for qualitative research in education. German Foundation for International Development (DSE)

Neill, J. (2007). Qualitative versus quantitative research: Key points in a classic debate. Accessed on 2 November 2009 from http://wilderdom.com/research/QualitatativeVersusQuantitativeResear.....

Newsom, J. T. (2007). A quick primer on exploratory factor analysis. Accessed on 23 December 2007 from http://www.ioa.pdx.edu/newsom/semclass/hoefa.doc

Nicol, D. J. & Boyle, J. T. (2003). Peer instruction versus class-wide discussion in large classes: a comparison of two interactions in the wired classroom. *Studies in Higher Education, 28(4), 457 – 473.*

Nitko, A. J. (1983). Educational tests and measurement: An Introduction. New York: Harcourt Brace Jovanovich.

Nunnaly, J. C. (1975). Psychrometric theory.25 years ago and now. *Educational Researcher, 4 (10), 7-14.*

O'Neil, T., Sireci, S. & Huff, K. F. (2002). Evaluating the content validity of a State mandated Science assessment across two successive administrations. (Available at School of Education, University of Massachusetts, 156 Hills House South Box 34140 Amherst, MA01003-4140).

Oosterhof, A. (2001). Classroom applications of educational measurement. New Jersey: Prentice-Hall.

OSET. (2008). Using active learning in the classroom. Accessed on 6 November 2008 from http://www.unm.edu/~OSET/

Palmer, T. M. (1978). In-service education: Intrinsic versus extrinsic motivation. In Rubin, L. (Ed.). (1978). The in-service education of teachers: Trends, processes and prescriptions (pp. 215 - 219). Boston: Allyn and Bacon.

Pedhazur, E.J. & Schmelkin, L. P. (1991). Measurement, design, and analysis: An integrated approach. New Jersey: Lawrence Erlbaum Associates.

POST. (2008). Research ethics in developing countries. Accessed on 7 June 2008 from www.parliament.uk/parliamentary_offices/post/

Poulis, J., Massen, C., Robens, E. & Gilbert, M. (1998). Physics lecturing with audience paced feedback. *American Journal of Physics 66(5), 439 - 441*.

Redish, E. F. (1994). Implications of cognitive studies for teaching physics. *American Journal of Physics 62(9), 796 – 803.*

Rennie, K. M. (1997). Exploratory and confirmatory factor rotation strategies in exploratory factor analysis. Accessed on 23 December 2007 from http://ericae.net/ft/tamu/

Riddle, E. M. and Dabbagh, N. (1999). Lev Vygotsky's social development theory. Accessed on 13 January 2007 from http://chd.gse.gmu.edu/

Rocco, T. S., Bliss, L. A., Gallagher, S. & Perez-Prado, A. (2003). Taking next step: Mixed methods research in organizational systems. *Information Technology and Performance Journal, 21(1), 19 - 29.*

Rubin, L. (Ed.). (1978). The in-service education of teachers: Trends, processes and prescriptions. Boston: Allyn and Bacon.

Russell, D. W. (2002). In search of underlying dimensions: The use (and abuse) of factor analysis. *Personality and Social Psychology Bulletin, 1629 – 1646.* Accessed on 7 December 2007 from http://psp.sagepub.com

Selemani-Mbewe, C. M. (2003). Knowledge, attitude and practice of classroom assessment: Implications on school based assessment in Malawi. Unpublished master's thesis, School of Education University of Massachusetts, Amherst.

Selemani-Meke, E. (2010). An assessment of the implementation of Continuing Professional Development programmes for Primary School teachers in Malawi: A case of Zomba Rural Education District. Unpublished PhD thesis, Faculty of Education at the University of Fort Hare. Available at Centre for Educational Research and Training, Chancellor College P.O. Box 280 Zomba Malawi.

Shepard, L. A. (1993). Evaluating test validity. *Review of Research in Education, 19, 405 – 450.*

Shepard, L. A. (1997). The Centrality of test use and consequences for testing. *Educational measurement: Issues and practices, 16(2), 5 - 8.*

Shultz, K. S., Riggs, M. L. & Kottke, J. L. (1998). The need for an evolving concept of validity in industrial and personnel psychology: Psychometric, legal and emerging issues. Current Psychology. Accessed on 4 April 2010 from http://www.accessmylibrary.com/article-1G1-54830469

Shumway, J. M. & Harden, R. M. (2003). The assessment of learning outcomes for the competent and reflective physician. *Medical teacher, 25 (6), 569 – 584.*

Sireci, S. G. (1998a). The construct of content validity. *Social Indicators Research, 45, 83 – 117.*

Sireci, S.G. (1998b). Gathering and analyzing content validity data. *Educational assessment 5(4), 299 – 321.*

Sireci, S. G. (2007). On validity theory and test validation. *Educational Researcher, 36(8), 477 - 841*

Sireci, S. G. & Geisinger, K. F. (1995). Using subject-matter experts to assess content representation: An MDS analysis. *Applied Psychological Measurement, 19(3), 241 – 255.*

Slavin, A. (2008). Peer instruction in the lecture setting. Accessed on 5 July 2008 from http://www.mcmaster.ca/cll/posped/pastissues/volume.1.no./

Southwest Educational Development Laboratory. (1995). Constructing knowledge in the classroom: Building an understanding of constructivism. Accessed on 19 July 2008 from http://www.sedl.org/scimath/compass/vo1no3/2.html

Sydenstricker-Neto, J. (2005). Research design and mixed methods approach: A Hands on experience. Accessed on 13 August 2005 from http://www.socialresearchmethods.net/tutorial/

Taleporos, E. (1998). Consequential validity: A practitioner's perspective. *Educational Measurement: Issues and practice, 17(2), 20 -23.*

Taylor, C.S. & Nolen, S.B. (1996). What does a psychrometrician's classroom look like?: Reframing classroom concepts in the context of learning. Accessed on 15 July 2007 from http://epaa.asu.edu/ojs/

Taylor, A. (2004). A brief introduction to factor analysis. Accessed on 3 August 2008 from http://www.psy.mq.edu.au/psystat/

Thorndike, R. M. (1997). Measurement and evaluation in Psychology and education. New Jersey: Prentice Hall.

Thornton, R. K. & Sokollof, D. R. (1998). Assessing students learning of Newton's Laws: The force and motion conceptual evaluation and the evaluation of active learning Laboratory and lecture curricula. *American Journal of Physics 66(4), 338 – 352.*

Training and Development Agency. (2008). What is continuing professional development? Accessed on 19 August 2008 from http://www.tda.gov.uk/

Tucker, M. L. & LaFleur E. K. (1991). Exploratory factor analysis: A review and illustration of five principal components decision methods for attitudinal data. Paper presented at the Annual Meeting of the Southwest Educational Research Association, San Antonio, Texas. Accessed on 13 November 2007 from http://www.ohio.ed/people/tuckerm1/Tucker%20vita.pdf

Tucker, L. R. & MacCallum, R. C. (1997). Exploratory factor analysis. Accessed on 13 November 2007 from http://medresearchconsult.com/factor.pdf

Tyler, R. W. (1931). A generalized technique for conducting achievement tests. *Educational Research Bulletin, 10(8), 199 – 208.*

Tyler, R. W. (1933). Assumptions involved in achievement test construction. *Educational Research Bulletin, 12(2), 29 – 36.*

University of Washington School of Medicine. (2008). Research ethics. Accessed on 7 June 2008 from http://depts.washington.edu/bioethx/topics/

Valentin, J. D. & Godfrey, J. R. (1996). The reliability and validity of tests constructed by Seychellois teachers. A paper presented at the 1996 joint conference organized by Educational Research Association (Singapore) and Australian Association for Research in Education November, 1996. Accessed on 16 June 2006 from http://www.aare.edu.au./96pap/godfj96269.txt

Vicky, R. N. (2009). Exploratory and confirmatory factor analysis. Accessed on 2 March 2010 from http://allnurses.com/nursing-articles/exploratory-confirmatory-factor-

Williams, P. (2008). Creating overwhelmingly successful learning programs for the(supposed) adults in your organization. Accessed on 21 July 2008 from http://www.articlesbase.com/authors/paula-williams

Worcester, D. A. (1934). On the validity of testing. *The School review, 42(7), 527 – 531.*

Yalow, E. S. & Popham, W. J. (1983). Content validity at the crossroads. *Educational Researcher, 12(8), 10-14*

Young, F. W. (1997). The nonlinear relationship of two variables: Nonlinear correlation coefficient. Accessed on 20 November 2009 from http://books.google.com/books*?*

Zhao, N., (2008). The minimum sample size in factor analysis. Accessed on 19 October 2008 from http://www.encorewiki.org/display/

APPENDICES

APPENDIX 3.1: SAMPLE PROFILE

Teacher	Class taught	Average class size	Other teaching	Qualification	Experience	Duties	Load (Normal 15 – 18)		
							School	Other	Total
1	Form 4 Physical science and Computer studies	120 50	Night school	DipEd, BSc(Comp)	4	Head of Night School, Boarding master and Head of Department	25	4	29
2	Form 2 Maths and Form4 Physical science	75	-	DipEd	2	AgHead of school, Patron CCAPSO	20	-	20
3	Form 3 Biology Form 4 Physical science	60	-	MSCE	2	Head of Department	38	-	38
4	Form 1 to Form 4 Physical science	80	Another school	DipEd	4	Sports mistress	36	6	42
5	Form 3 to Form 4 Physical science	130	Night school	BSc	2	Timetable master, Patron of HIV/AIDS and Wildlife clubs	25	9	34
7	-	-	-	BSc	7	-	-	-	-
8	Form 2 Agriculture Form 4 Maths and Form 1 to Form 4 Physical science	75	Night and Private school	BSc	2	Form teacher	17	13	40
9	Form 1, Form 3 to Form 4 Physical science	80	Night school	DipArch	12	Exam Committee member	28	9	37
10	Form 2 Maths Form 4 Physical science	80	-	BScEng	5	Head of Department	22	-	22

Teacher	Class taught	Average class size	Other teaching	Qualification	Experience	Duties	Load (Normal 15 – 18)		
11	Form 2 and Form 4 Physical science	120	Night school	DipEd	1	Library, Head of Department, Patron YCS	18	3	21
12	Form 1to Form 4 Maths and Physical science	70	Night school	DipEng	4	Head of Department Deputy Head	25	10	35
13	Form 2 Maths and Form 4 Physical science	120	Private schools	DipEd	26	Patron Wildlife	20	5	25
14	Form 3 Maths and Form 4 Physical science	80	Night school	BSc	2	Sports and Entertainment master	22	15	37
15	Form 2 and Form 4 Physical science	90	Night school	DipEd	12	Head of Department	20	5	25
16	Form 1 to Form 4 Physical science	70	Night school	BSc	26	Head of Department Form master, Coach	24	2	26
17	Form 4 Physical science	70	-	DipEd	11	Head of Department	15	-	15
18	Form 1 Agriculture Form 2 Maths and Form 4 Physical science	-	Night school	DipED	20	Head of Department	30	9	39

APPENDIX 3.2: TEACHERS' TESTS

Teacher 1/ test 1
End of Term 1 test
MSCE Physical science

Instruction
Answer all questions in the space provided.

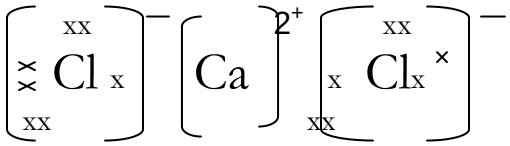
1. Study Table 1 which shows particles found in the atoms of 4 elements below.

Element	Protons	Neutrons	Electrons	Mass number
Hydrogen (H)	1	-	-	1
Carbon (C)	-	-	6	12
Nitrogen (N)	7	12	-	-
Sodium (Na)	-		11	-

Table 1: Atomic particles

a) Complete the table by filling in the missing numbers. (8)
b) (i) Which element in the table will easily form an ionic compound? (1)
(ii) Give reasons for your answer. (2)
c) Work out the molecular mass of methane (CH_4). (3)
d) What kind of chemical bonds are involved in methane? (1)
e) Explain your answer. (3)

2. a) The dot and cross diagram of calcium chloride is shown below.



(i) Write down the chemical formula of calcium chloride. (1)
(ii) Explain the meaning of 2+ on the Ca atom. (2)

b) Table 2 shows elements represented by letters Q, R, L, X, W, Y and Z in the same periodic table. Study it and answer questions that follow:

Group	I	II	III	IV	V	VI	VII	VIII
Elements	Q	R	L	M	X	W	Y	Z

(i) Write the formula of a charged atom. (1)
(ii) Give a letter of an element in the table
1. that would not react with another element. (1)
2. which belongs to the halogen

(iii) Why is it that the element you have chosen in (ii) 1 above would not react with another? (2)
(iv) Is it possible that 'Z' can be Helium. Helium has 2 electrons in Its outermost shell and yet it is placed in group 8. Explain why

this is not by mistake. (2)

3. a) Table 3 below shows results in an experiment to verify a gas law. Study it and answer questions that follow.

Volume (cm^3)	10	12	14	16	18
Elements	Q	R	L	M	X

(i) Plot a suitable graph on the graph paper provided to show the relation between pressure and volume. (6)

(ii) What relation is being demonstrated by this graph? (1)

(iii) What gas law is being stated in (ii) above? (1)

b) Study Figure 1 below and answer questions that follow.

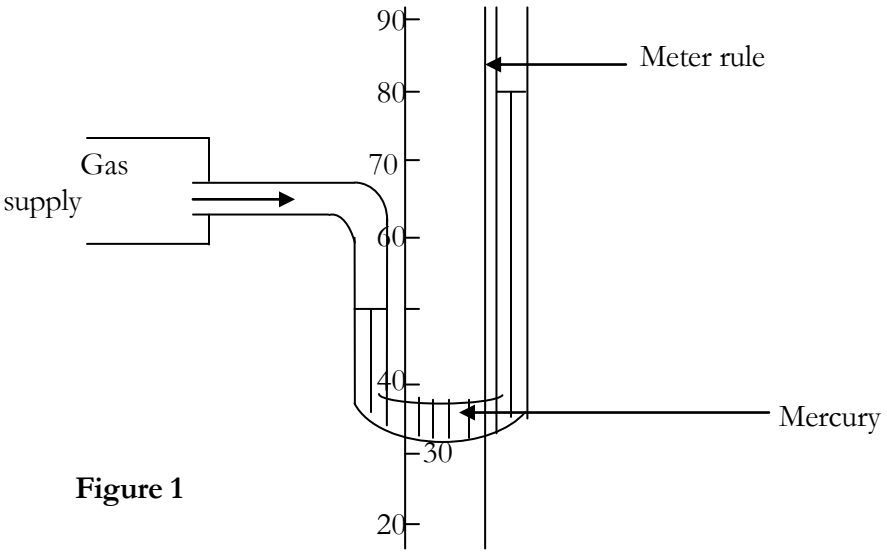


Figure 1

(i) Name the instrument above. (1)

(ii) What is the pressure difference in mmHg? (1)

(iv) If the atmospheric pressure is 755mmHg what is the pressure of gas supply? Show your working clearly. (2)

(v) What is the difference between the instrument in fig. 1 above and a Constant Gas Volume Thermometer? (2)

c) Explain why dams are made thicker at the bottom and left thinner on top? (2)

4. a) Fig. 2 is a speed-time graph for journey made by Shire Bus.

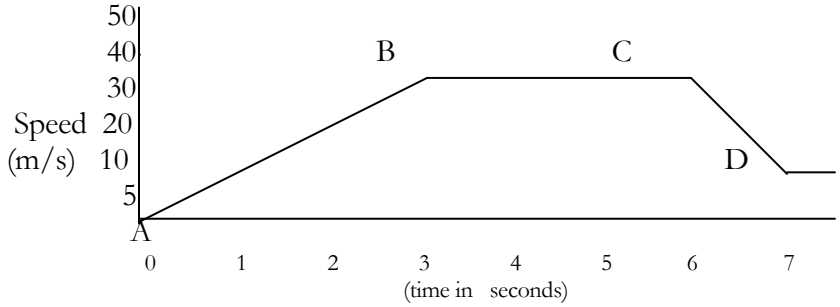


Figure 2 Speed-time graph

(i) Describe the motion of the bus between A and D. (3)
(ii) Calculate the acceleration of the bus between A and B. (2)
(iii) Calculate the total distance traveled between A and C. (3)

b) A piece of a card of 20 cm diameter and a coin of equal masses were dropped from a height of 2m above the ground.
(i) Assuming that air resistance is not negligible, which of the two would reach the ground first?
(ii) Explain your answer in b(i) above.
(iii) State the three forces which act on each object as it falls. (3)
(iv) Which forces would remain constant as the object falls? (2)

5. With the aid of relevant examples and diagrams explain the difference between
(i) Covalent bonding and ionic bonding. (5)
(ii) Isomerism and conformation. (5)
6. A laboratory technician has one litre of 2M hydrochloric solution. Describe how she would prepare a 250cm^3 volume of 0.2M hydrochloric acid from the 2M solution.

Teacher 1 / test 2
Parallel test to end of Term 1 test
MSCE Physical science

1. a. Define
(i) absolute temperature (1)
(ii) gas pressure (1)
b. With aid of a diagram (where possible) explain how knowledge of solid expansion and contraction can be used in each of the following:
(i) riveting metal plates (3)
(ii) shrinking fitting (3)
c. A cyclist checks his bicycle tire at start of a journey and finds that it has a pressure of 5000 kpa. On reaching his destination, it was found to be 2000 kpa. If the temperature on reaching his destination was 10°C what was the temperature on start? (Assume that the volume was constant.) (3)

2. a. (i) Name one atomic particle that has no charge. (1)
(ii) An element, sodium, has 11 electrons and its mass number is 23. Draw a full electronic shell configuration using circles and label parts. (3)
(iii) An element X has an atomic number of 35.
1. In which period would you place it in the periodic table? (3)
2. Explain 2 chemical properties the element X would have. (2)

(iv) Atoms of Nitrogen combine to form N_2 molecules.
1. What type of non-polar covalent bond do these form? (1)
2. Draw an electron dot and cross diagram to show how these atoms combine to form a molecule.
3. Based on your drawing above, how would the molecule of N2 above differ from that of NACl in terms of :
- Bond formation.

- Melting and boiling point.

(v) Work out the formula of the compound formed when

1. Hydrogen combines with Sulphur. (2)
2. Berilium combines with Iodine. (2)

b. (i) Which of the following compounds are polar covalent? H_2O, CO_2, MgO (1)

(ii) Why is a metallic bond described as cations in a sea of electrons? (1)

c. Study Table 1 of metals and their properties and then answer the questions that follow?

	Metals	**Properties**
1.	Hard steel	Tough and brittle
2	Stainless steel	Tough, does not corrode
3	Solder (70% lead and 30% tin	
4	Brass (copper + Zinc)	Silvery, attractive

(iii) Why is the property of stainless steel good for making it to be used for kitchen utensils?

(iv) What is the common use for Brass?

(v) Explain the effect of the strength of intermolecular forces (IMF) on melting and boiling points of the halogens. (1)

b. What halogen is used in the following substances?

(ii) Common salt (1)

(iii) Colgate (1)

(iv) PVC plastic (1)

(v) Photographic film (1)

d. (i) Explain how halogen compounds contribute to environmental hazards. (4)

(ii) State the three main sources of Sulphur. (3)

(iii) Explain the differences between Rhombic and Mono-clinic Sulphur. (2)

(iv) Other than the manufacturing of Sulphuric acid, state the five uses of Sulphur. (5)

(v) How important are Sulphates in

1. Crop sciences. (1)
2. Medical technology (1)

3. a. (i) What is a chemical reaction?

(ii) From the above definition, explain the implication on the status of reactants and products in a chemical reaction.

(iii) How many particles are there in 1 mole of a given substance? (1)

(iv) Define an empirical formula of a compound. (1)

b. (i) In an experiment an unknown organic compound was found to contain 0.12 g of carbon and 0.02g of hydrogen. Calculate the empirical formula for the compound. (3)

(ii) If the mass of a molecular formula of in 'b(i)' above was found to be 56g, work out the molecular formula for the compound. (RAM: C = 12, H = 1) (3)

c. (i) State 3 ways through which concentration of a substance can be expressed. (3)

(ii) 14.6g of HCl were used in making up a solution of volume 2 litres. 30 centilitres of this solution neutralized 25cm^3 of sodium carbonate

solution. Work out:

1. Molarity of HCl acid.
2. Concentration of sodium carbonate solution in moles per cubic

d. A chemist was given a concentrated 4 litres bottle HCl of an unknown Molarity. Upon drawing 50 cm^3 from the bottle, he prepared 500$cm^3$0.5M solution. Calculate the Molarity of the highly contentrated 4 litre bottle. (4)

e. Explain what is meant by Molar volume of gas.

f. A pure gas of hydrogen (H_2) was found to occupy 72 cm^3. Another gas carbon dioxide (CO_2) was found to occupy 72cm^3. Calculate number of moles of:

(i) H_2 contained in 72cm^3. (3)

(ii) CO_2 in 72cm^3. (1)

(iii) Explain the relationship of your answer in b(i) and b(ii) above. (2)

g. With the aid of a diagram, explain what each of the following means:

(i) Activation energy. (4)

(ii) Exothermic reaction. (4)

(iii) Endothermic reaction (4)

h. Explain with examples of reactions why:

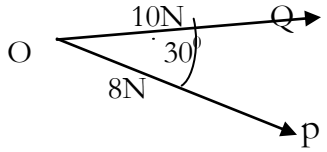
(i) Bond formation is endothermic. (3)

(ii) Bond breaking is exothermic. (3)

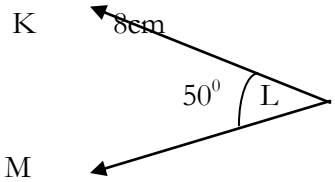
4. a. What is the difference between a vector quantity and a scalar one? (1)

b. Find a resultant force for each of the pairs on this paper.

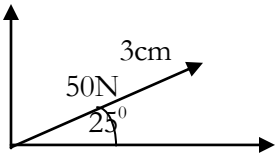
(i) By triangle rule



(ii) By parallelogram rule (3)



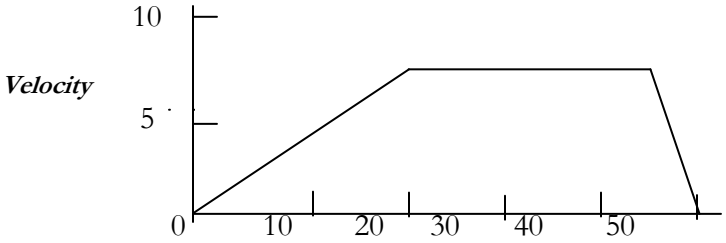
(iii) Resolve the following vector into a horizontal and vertical component, hence calculate the magnitude of each component. (4)



(iv) Distinguish between

1. distance and displacement. (1)
2. speed and vector. (1)
3. zero acceleration and zero velocity. (1)

c. Study the following velocity – time graph and answer the questions that follow:



Find:

(i) the maximum velocity. (1)

(ii) the acceleration during the first part of the journey. (2)

(iii) the total distance traveled. (3)

d. State Newton's Third law of motion. (1)

e. Define terminal velocity. (1)

f. A small rocket of mass 200 kg takes off with an engine upward force of 300N. Calculate:

(i) The resultant force that brings about the upwards motion. (3)

(ii) The acceleration of the rocket. (3)

5. a and b *disqualified by the examiner.*

c. Name two catalysts which act as an oxidizing agent in a reaction to form a carboxylic acid from an alcohol. (2)

d. A certain carboxylic acid has a total of 10 hydrogen atoms.

(i) Write down the molecular formula of this carboxylic acid. (1)

(ii) Draw the structure of the carboxylic acid in d(i) above. (2)

e. State the three natural sources of carboxylic acids. (3)

f. Ethanoic acid was reacted with ethanol in the presence of drops of concentrated (sulphuric acid) H_2SO_4.

(i) State the two products of this reaction. (2)

(ii) Write down:

1. Word equation for this reaction. (1)
2. Chemical equation for the reaction. (2)

a. What name is given to the process above? (1)

(iii) You are given unlabelled bottles with the following: hexanol butane, pentene and propanoic acid. You are also provided with labeled bottles of sodium, distilled water and bromine solution. Using materials provided, describe 3 basic steps you would do to identify the compounds in the bottles. Summarise your expectations using a flow diagram. (6)

6. a. What are isomers? (1)

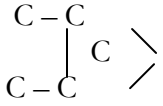
b. Study the following structures and answer the questions that follow:

A. C = C – C – C – C

B. C = C – C – C

C. C – C – C = C

D.

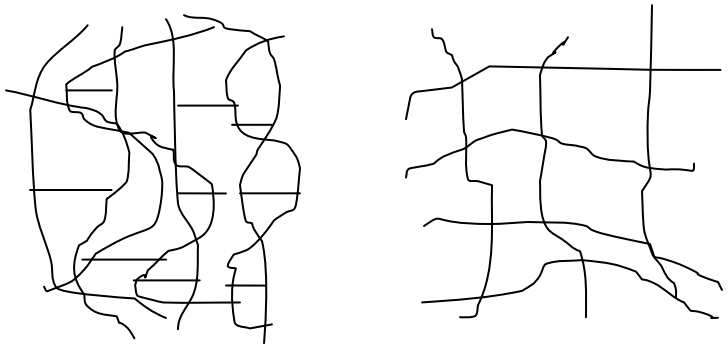


(i) …… is a conformation to ……. (1)
(ii) ……... is an isomer to …… (1)

c. Hydrocarbon of longest carbon chain of 5 has a C_2 group attached on carbon atom Number 2 and 2CH groups attached on carbon atom number 3. Assuming that it has no double bond,
(i) Draw the structure of the hydrocarbon. (3)
(ii) Name the hydrocarbon. (1)
(iii) Write down its condensed structural formula. (1)

d. Study the following chemical equations and answer the questions that follow.
(i) $H_2NCH_2COOH + H_2NCHOH_3COOH \longrightarrow H_2NCH_2CONHCH_3COOH + H_2O$
(ii) $HO(CH_2)_2)OH + HOOC(C_2H_2)COOH \longrightarrow HO(CH_2)_2OCO(C_6H_4)COOH + H_2O$
1. What type of polymerization is shown in the equations above? (1)
2. Give reasons for your answer in (ii) 1 above. (1)
3. Name linking blocks in the equation. (2)

e. Which equation represents protein synthesis? (1)
f. Give a reason for your answer in (e) above. (1)
g. Explain two differences that lead to production of polythene and the processes represented by the equations above. (2)
h. Study the polymer chains below and answer questions that follow.

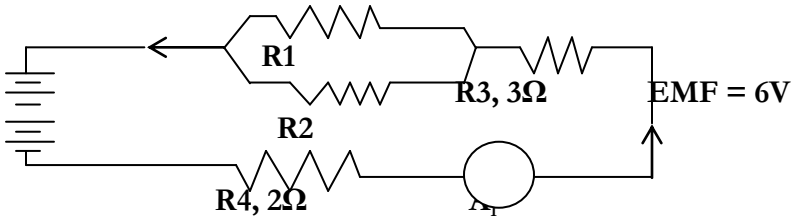


A B

(i) Which set of polymers softens when heated? (1)
(ii) Explain your answer in (i) above. (2)
(iii) Biodegradable and hydro-degradable plastics are not yet common in our current technology. Explain ways of disposing plastics made by the current technologies. (3)

7. How is the knowledge of electrostatics used in everyday life in each of the following?
1. Capacitors (3)
2. Photocopiers (3)
3. Precipitation (3)

Study the following circuit diagram and answer the questions that follow. Each cell is a standard cell used in torches.



Calculate:

(i) the total EMF put in by the cells. (1)
(ii) the potential difference across R4 if A1 reads 2 Amperes. (3)
(iii) the potential difference across R3. (3)
(iv) the current at R1. (3)
(v) the current at R2. (3)
(vi) the total resistance of R1 and R2. (3)

8. Discuss three practical pressure applications in life based on each of the states of matter which have no definite shape. (8)

10. Describe the procedure you would follow to prepare ethanol using the indigenous method. (8)

Note: More of the teachers' tests are on the CD.

APPENDIX 3.3: BASELINE INFORMATION ABOUT THE POPULATION

Note: This form should be completed by a teacher appointed to teach a Form 3 Physical Science class in the 2006 academic year (1st Term starting in January 2006)

Information about the school

1. Name of the school _____ Phone number _____
2. Classification (Underline what applies):
   Boys Girls Co-ed Private Full Boarding Partly Boarding
3. Main selection mechanism of learners (Tick):
   Government selection Entrance examinations Verbal/written applications
4. Qualification of Head_____ Experience as Head_____
5. Expected number of 2006 Form 3 Physical Science streams (classes) ____
6. Current level of resources (chemicals and equipment) for teaching 2006 Form 3 and 2007 Form 4 Physical Science in the school (Be fair and tick against one of the ratings below)
   - None
   - Inadequate and can manage only for some of the MSCE Physical Science topics
   - Inadequate but can manage for the MSCE Physical Science course
   - Just adequate for the Physical science course
   - Comfortable
7. What facilities does the school use for printing class tests?
   _____

Information about the teacher

8. Full name of teacher _____Sex _____ Phone____
9. Teacher's highest qualification_____
10. College attended _____ Year of graduation _
11. Teaching experience in MSCE Physical Science (Complete Table below)

Year	Form	School	Year	Form	School

12. Number of Form 3 Physical Science streams (classes) to teach in 2006. _____
13. Estimated overall teaching load per week in 2006 academic year. _____
14. Estimated number of 2006 Physical Science teaching load per week. _____
15. Tick the following MANEB activities in the table below in which you participate.

	Item writing	Item critiquing	Item editing	Moderation	Marking
Tick					
Period					
Subject					

APPENDIX 3.4: EVALUATION OF WORKSHOP CONTENT

A. Personal information

Name: _____Sex_____
Tel/Cell No. _____
School: _____Class size _____
Qualifications: _____College _____
Teaching experience: General _____ P/ science _____MSCE P/Science __
Involvement in MANEB activities:
a. Test development _____years Subject _____
b. Marking _____Years Subject _____

Questionnaire

A workshop on classroom test construction has been designed to draw upon your teaching experience in order to build a common understanding of different concepts and principles involved. Which of the following areas do you recommend for review in the content of such a test construction workshop? Tick in the box on the right of an area to show degree of recommendation.

Key: 1. Least to 3. Most

	1	2	3
1. Definition of a test	☐	☐	☐
2. Description of a test	☐	☐	☐
3. Purpose of classroom tests	☐	☐	☐
4. Coverage of classroom tests	☐	☐	☐
5. Writing good test items	☐	☐	☐
6. Determining order of test items	☐	☐	☐
7. Quality of classroom tests	☐	☐	☐
8. Assembling items for classroom tests	☐	☐	☐
9. Preparing a marking scheme	☐	☐	☐
10. Assessing performance of items in a classroom test	☐	☐	☐

Thank you for responding to the questionnaire

APPENDIX 3.5: WORKSHOP EVALUATION

Questionnaire

Considering your experience in this workshop on construction of classroom tests which of the following areas:

1. ***Did you find useful*?** Tick in the box on the right of an area to show the extent you found it useful.

Key: 1. Least to 3. Most

	1	2	3
1. Description/definition of a test	☐	☐	☐
2. Purpose of classroom tests	☐	☐	☐
3. Coverage of classroom tests	☐	☐	☐
4. Writing good test items	☐	☐	☐
5. Writing higher order test items	☐	☐	☐
6. Quality of classroom tests	☐	☐	☐
7. Balancing classroom tests	☐	☐	☐
8. Preparing a marking scheme	☐	☐	☐
9. Assessing performance of items in a classroom test	☐	☐	☐
10. The workshop as a whole.	☐	☐	☐

2. ***Did youfind relevant?*** Tick in the box on the right of an area to show the extent you found it to be relevant.

Key: 1. Least to 3. Most

	1	2	3
1. Definition of a test	☐	☐	☐
2. Description of a test	☐	☐	☐
3. Purpose of classroom tests	☐	☐	☐
4. Coverage of classroom tests	☐	☐	☐
5. Writing good test items	☐	☐	☐
6. Writing higher order test items	☐	☐	☐
7. Quality of classroom tests	☐	☐	☐
8. Balancing classroom tests	☐	☐	☐
9. Preparing a marking scheme	☐	☐	☐
10. Assessing performance of items in a classroom test	☐	☐	☐
11. The workshop as a whole	☐	☐	☐

3. ***Do you understand better after the workshop?*** Tick in the box on the right of an area to show the extent to which you understand them better.

Key: 1. Same as before 2. Slightly better 3. Better 4. Much better

	1	2	3	4
1. Definition of a test	☐	☐	☐	☐
2. Description of a test	☐	☐	☐	☐
3. Purpose of classroom tests	☐	☐	☐	☐
4. Coverage of classroom tests	☐	☐	☐	☐
5. Writing good test items	☐	☐	☐	☐
6. Writing higher order test items	☐	☐	☐	☐
7. Quality of classroom tests	☐	☐	☐	☐
8. Balancing classroom tests	☐	☐	☐	☐
9. Preparing a marking scheme	☐	☐	☐	☐
10. Assessing performance of items in a classroom test	☐	☐	☐	☐
11. The workshop as a whole	☐	☐	☐	☐

Thank you for responding to the questionnaire

APPENDIX 3.6: FORM FOR CODING IN ITEM ANALYSIS

The form used for coding scores of candidates from scripts for item analysis was as given below.

Test: _____

Maximum mark: _____

Sch. No.	Stud. No.	X_T	X_Q	Sch. No.	Stud. No.	X_T	X_Q	Sch. No.	Stud. No.	X_T	X_Q

An adaptation from MANEB

Key: Sch. No. – School number which was also a teacher's number

Student No. – number of a learner in class

X_T - learners' total score on the test

X_Q – learners' score on the item

APPENDIX 3.7: ITEM REVIEW FORM

The item review form was a form which Subject Matter Experts used to rate the items for relevance and representativeness. It had the details given below.

Instruction: Place in the appropriate cell of a topic of the test a 1 or 2 or 3 or 4 or 5 toshow the extent to which an item measures the topic. Do the same for a cognitive level to show the extent to which the item measures the topic to that level of cognitive ability. 1 is least and 5 most in both cases.

Thank you for your support.

Item number	Topics						Cognitive level		
							Recall	Comprehension	Higher order

An adaptation from Sireci(1998b)

APPENDIX 3.8: REQUEST FOR ADMINISTRATION OF QUESTIONAIRES

THE MALAWI NATIONAL EXAMINATIONS BOARD

P. O. BOX 191, ZOMBA, MALAWI, TEL. (265) 525 277, CABLE: MANEB, ZOMBA TELEX: 44428, MANEB MI, FAX: (265) 525 351
e-mail: maneb@chirunga.sdnp.org.mw

EXECUTIVE DIRECTOR
Mr M. W. Matemba

All communications should be addressed to:
The Executive Director

In reply please quote:

Our Ref.:

Your Ref.:

9th August 2005

The Head
Ben-nevis Secondary School
P.O. Box 90346
Blantyre

Dear Sir/Madam,

BASELINE INFORMATION ABOUT 2006 FORM 3 PHYSICAL SCIENCE TEACHERS

I am interested in following up the teaching of Physical science in secondary schools particularly at MSCE level. The purpose for writing you is to find out whether your school will offer Physical science to a Form 3 class in 2006 academic year and the teacher who will take this class.

A form has been attached which I should kindly ask if the teacher identified to teach 2006 Form 3 Physical science class could complete. I will be most grateful to you if the form can be sent back to me by end September 2005. The form should be sent back even if the school will not have a 2006 Form 3 Physical science class.

Your support and quick attention will be deeply appreciated.

Yours faithfully

Mike Nkhoma

Attached .....

CC: Executive Director
The Malawi National Examinations Board

APPENDIX 3.9: REQUEST TO INVOLVE SCHOOLS IN THE RESEARCH

C/O The Malawi national Examinations Board
P.O. Box 191
Zomba

8th August 2005

Funded
16-08-05

Through: The Executive Director,
Malawi National Examinations Board
P.O. Box 191
Zomba

To: The Secretary for Education
Ministry of Education and Culture
P/Bag 328, Lilongwe 3
Fax No. 01789662

Attention Mrs Kabuye

Dear Sir,

REQUEST FOR PERMISSION TO COMMUNICATE WITH SECONDARY SCHOOLS

I am a student of doctoral studies with the University of Malawi. I am interested in the teaching of Physical science in secondary school particularly at MSCE level. I would, therefore, be interested to trace schools that will offer Physical science to a Form 3 class in 2006 and teachers who would teach it.

The purpose for writing you is to seek your permission to communicate with Heads in your secondary schools in the Southern Region to get more information about the schools and the Physical science teachers. To this effect a questionnaire whose sample is attached will be sent to the schools once approval is given.

The information sought will serve as baseline information to help me select a sample of schools and teachers for my research study. It is expected that the selected teachers will undergo a kind of an indepth in-service training in principles of assessment which will help them carry out school based assessments effectively. The training tentatively is scheduled for November or December when schools are on holidays. Thereafter, their progress in assessing their pupils up to the period they write 2007 MSCE will be monitored by the researcher. The objective is to find out whether valid school based assessments in Physical science would be an effective alternative to the observed low performance of the candidates in this subject.

I will be most grateful for your kind consideration.

Yours faithfully,

Mike Nkhoma

Attached……

APPENDIX 3.10: APPROVAL TO INVOLVE THE SCHOOLS

65

THE MALAWI NATIONAL EXAMINATION BOARD

INTER-OFFICE-MEMORANDUM

To: DED **Ref:**

Through: Director RTD **Date:** 29/08/05

From: SRTDO

Subject: Communication with schools for sampling

I was following up with Director EMAS on the request submitted to MOE for permission to send a questionnaire to their schools. Her verbal response was that the questionnaires could be sent to the schools to avoid delays. Since she is out of office she was hopeful that the PS was going to approve it. In case of queries she would be of reference.

Since its getting late to collect information from schools to assist with proper sampling, I am not sure whether it is in order to go for her word or to try with Division Managers, South East, South West and Shire Highlands where these schools are.

I will be very grateful for Management's guidance.

M Nkhoma

II

ED

Sir.

As per your directive on the above issue, I've advised Mr. Nkhoma to send questionnaires to schools.

III

SRTDO 30/8/05

Mike follow the advice of DED. I trust Director EMAS she does what she says Mk 06.09.05

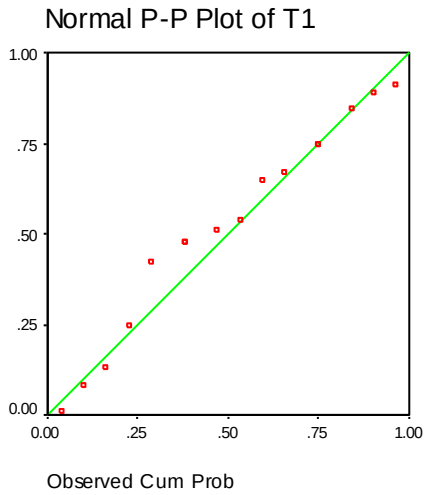
APPENDICES 4.1: COMPUTATION

APPENDIX 4.11

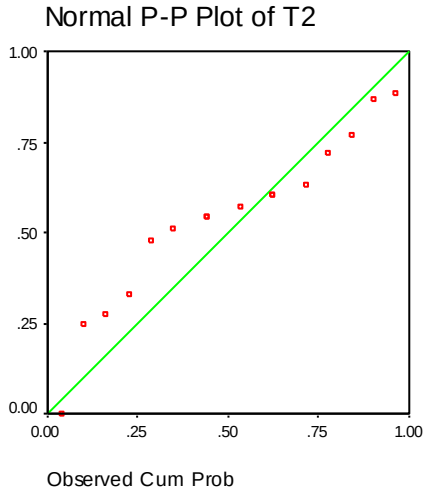
P-P PLOT FOR NORMALITY OF DISTRIBUTION

a. Item discrimination

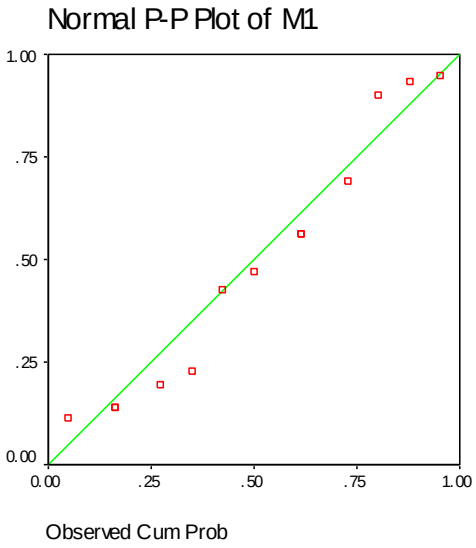
(i) Percentage of good and excellent items combined of T1



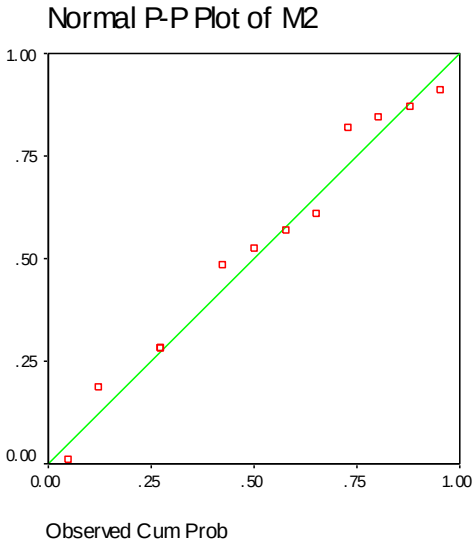
(ii) Percentage of good and excellent items combined of T2



(iii) Percentage of good and excellent items combined of M1

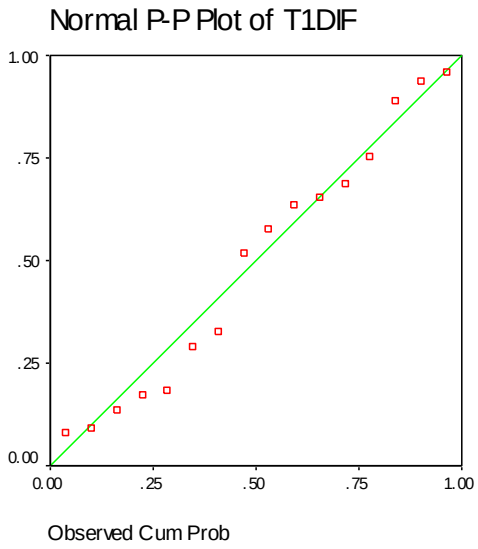


(iv) Percentage of good and excellent items combined of M2



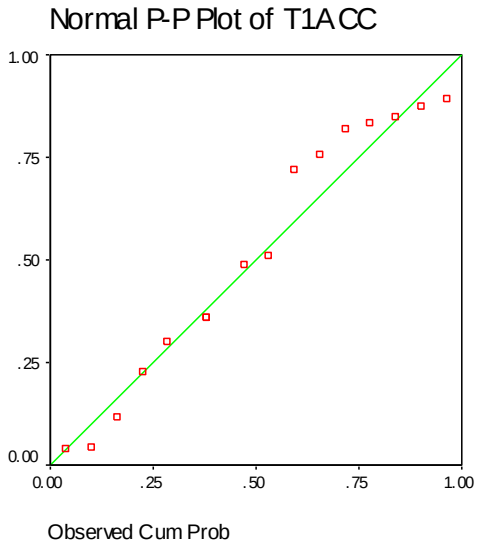
b. Item difficulty

(i)Difficult items of T1



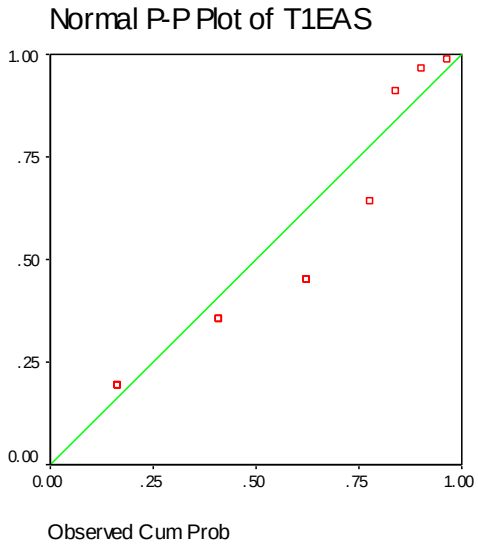
Key: T1DIF – Difficult items of T1

(ii) Acceptable items of T1



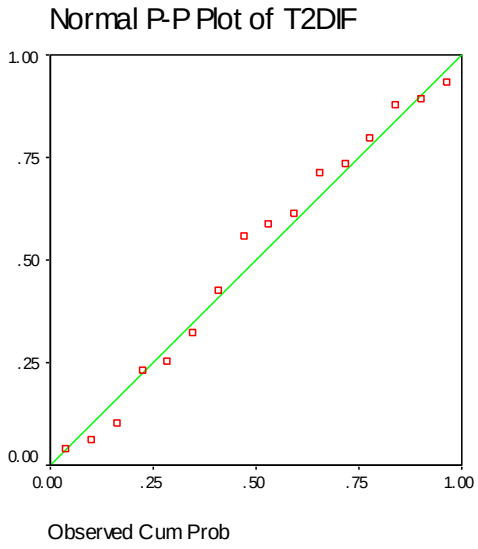
Key: T1ACC - Acceptable items of T1

(iii) Easy items of T1



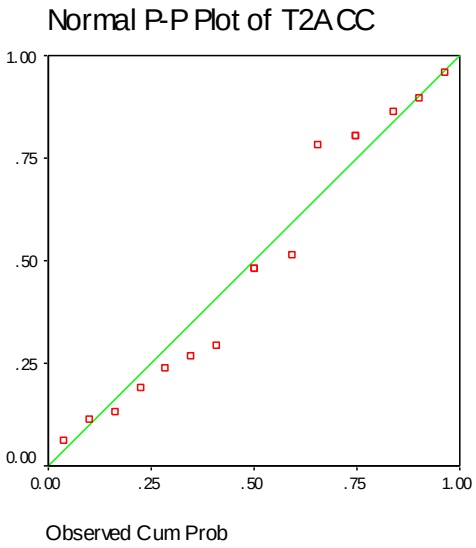
Key: T1EAS – Easy items of T1

(iv) Difficult items of T2



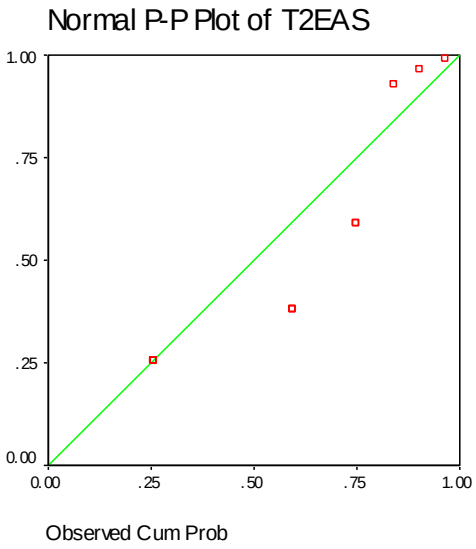
Key: T2DIF – Difficult items of T2

(v) Acceptable items of T2



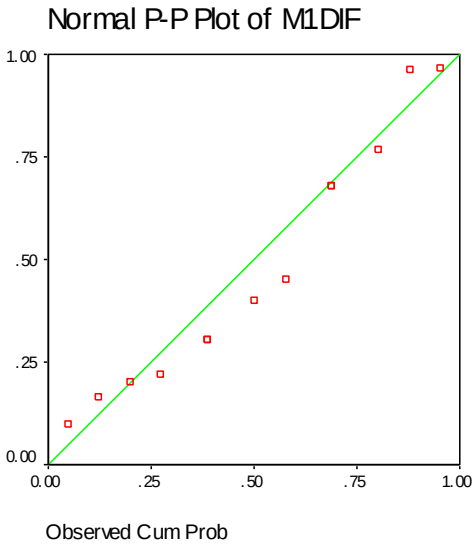
Key: T2ACC – Acceptable items of T2

(vi) Easy items of T2



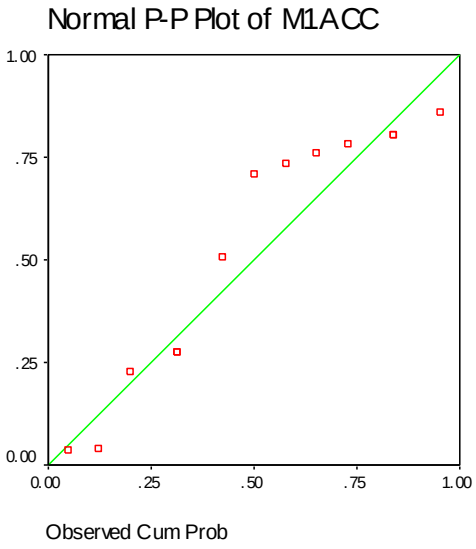
Key: T2EAS – Easy items of T2

(vii) Difficult items of M1



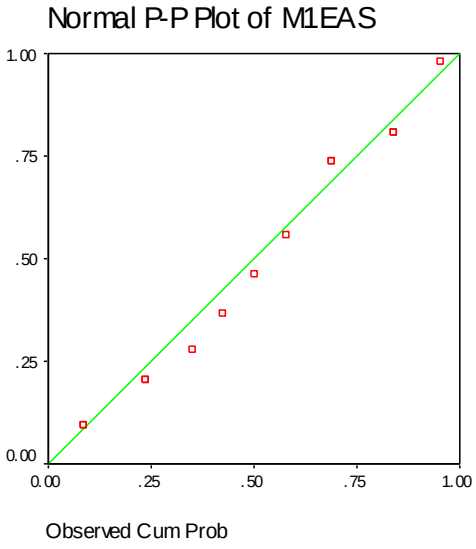
Key: M1DIF – Difficult items of M1

(viii) Acceptable items of M1



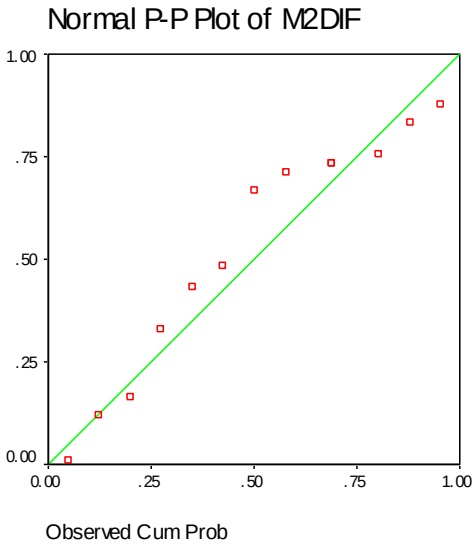
Key: M1ACC – Acceptable items of T1

(ix)Easy items of M1



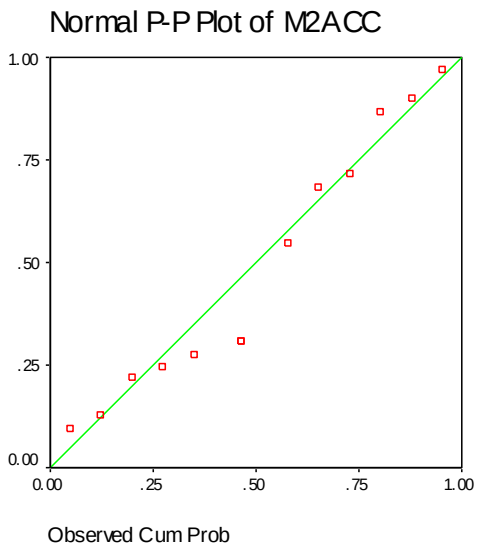
Key: M1EAS – Easy items of M1

(x) Difficult items of M2



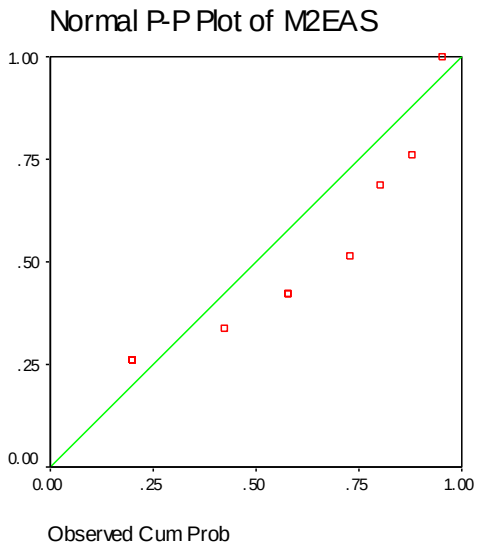
Key: M2DIF – Difficult items of M2

(xi) Acceptable items of M2



Key: M2ACC – Acceptable items of M2

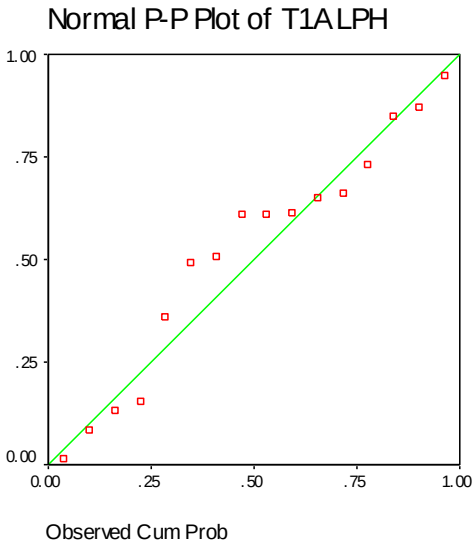
(xii) Easy items of M2



Key: M2ACC – Acceptable items of M2

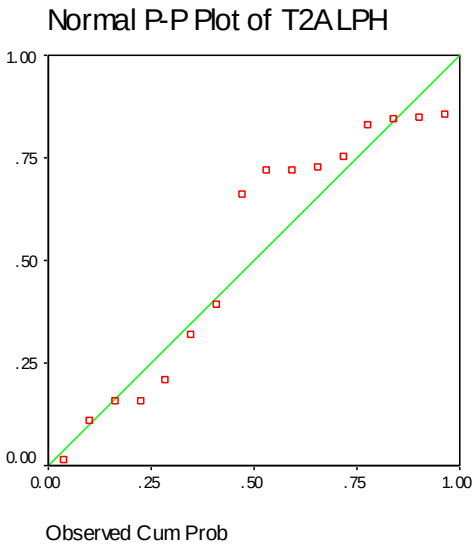
c. Reliability of pre-tests and post-tests

(i) Alpha reliability coefficients of T1



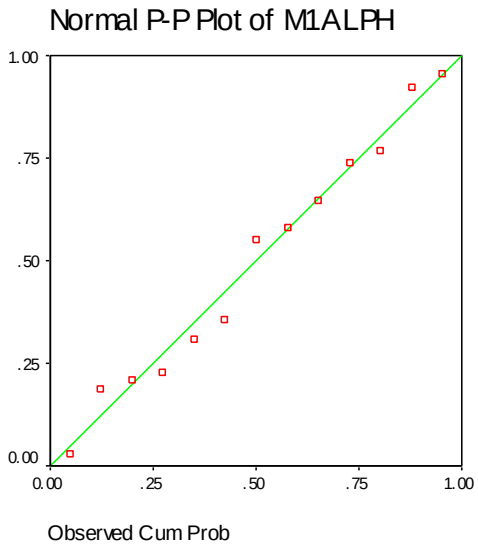
Key: T1ALPH – Alpha coefficients of T1

(ii) Alpha reliability coefficients of T2



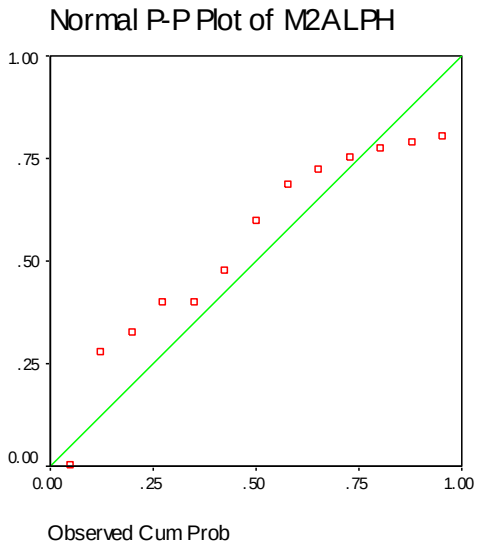
Key: T2ALPH – Alpha coefficients of T2

(iii) Alpha reliability coefficients of M1



Key: M1ALPH – Alpha coefficient of M1

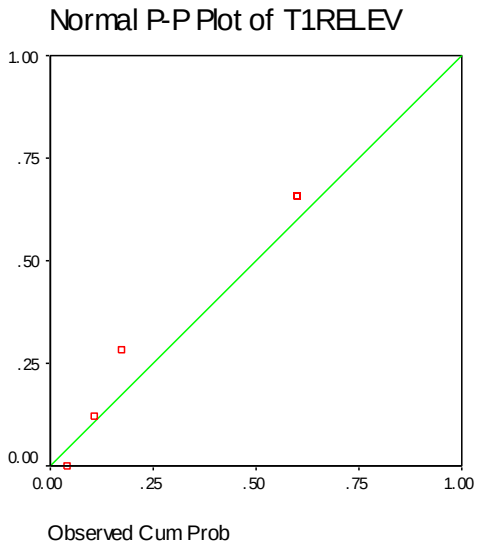
d. Alpha reliability coefficient of M2



Key: M2ALPH – Alpha coefficient of M2

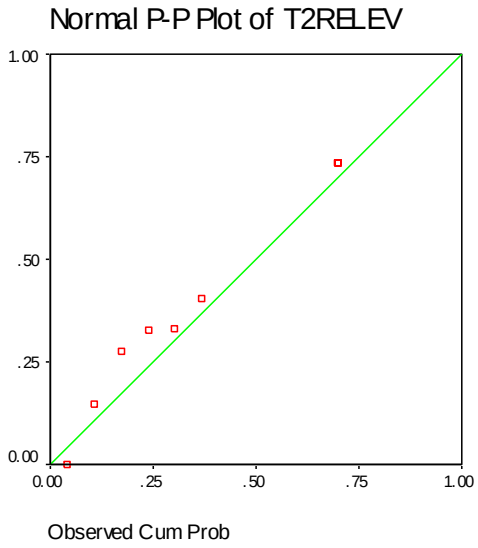
e. Item relevance rating of pre-tests and post-tests

(i) Item relevance ratings of T1



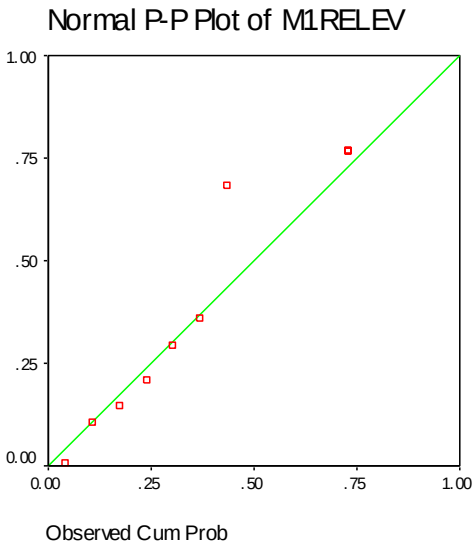
Key: T1RELEV – Item relevance ratings of T1

(ii) Item relevance ratings of T2



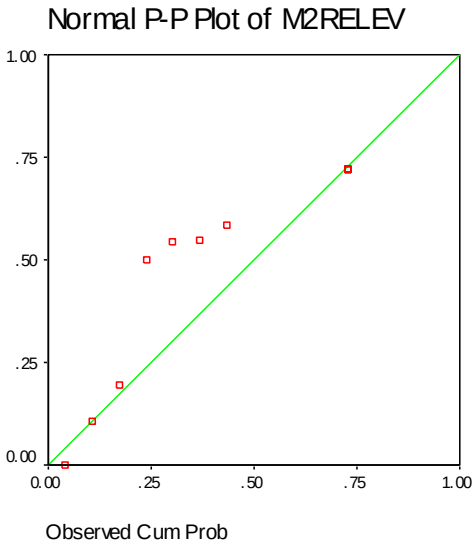
Key: T2RELEV – Item relevance ratings of T2

(iii) Item relevance ratings of M2



Key: M1RELEV – Item relevance ratings of M1

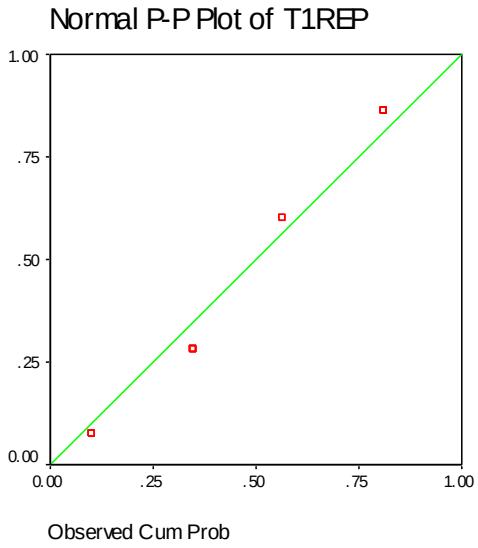
(iv) Item relevance ratings of M2



Key: M2RELEV – Item relevance ratings of M2

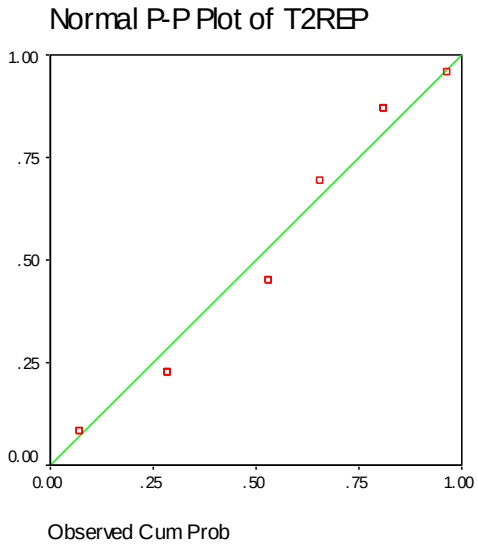
e. Rating for item representativeness

(i) Item representativeness ratings of T1



T1

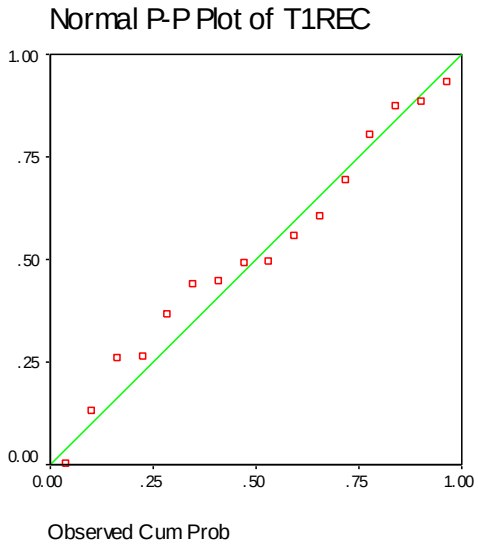
(ii) Item representativeness ratings of T2



T2

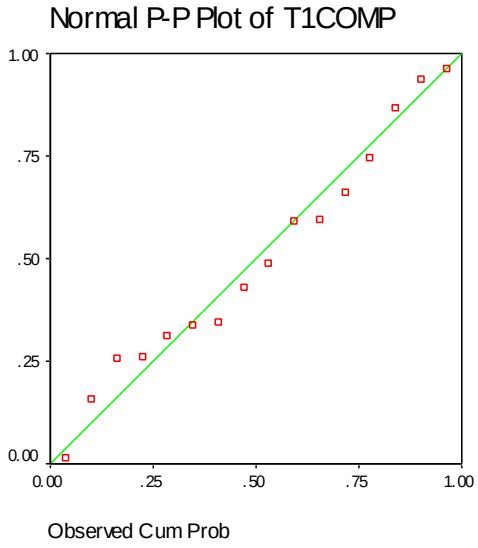
f. Cognitive rating of items of teachers' tests

(i) Item cognitive ratings of T1 at recall level



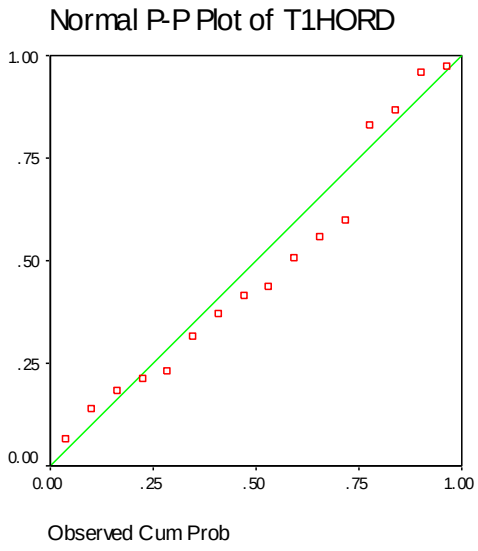
Key: T1REC – T1 at recall level

(ii) Item cognitive ratings of T1 at comprehension level



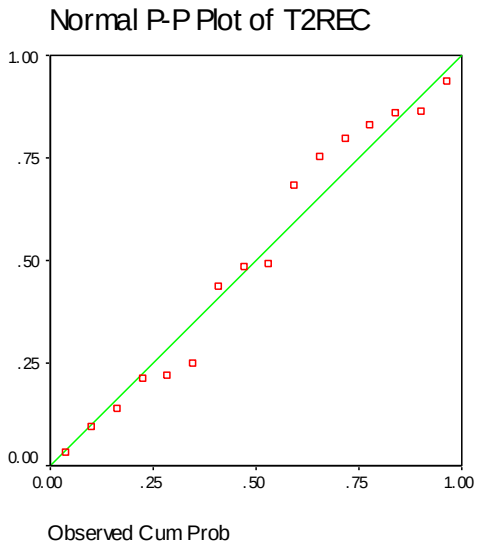
Key: T1COMP – T1 at comprehension level

(iii) Item cognitive ratings of T1 at higher order level



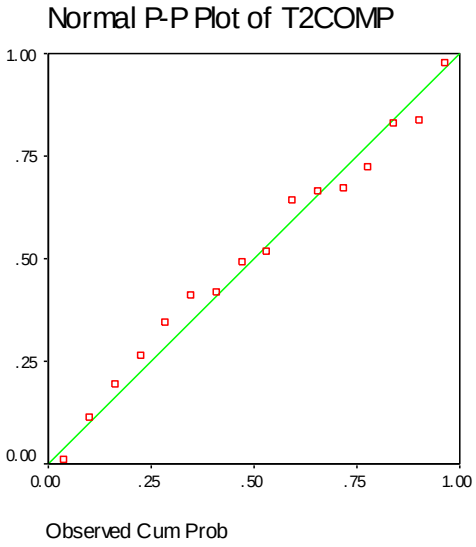
Key: T1HORD – T1 at higher order level

(iv) Item cognitive ratings of T2 at recall level



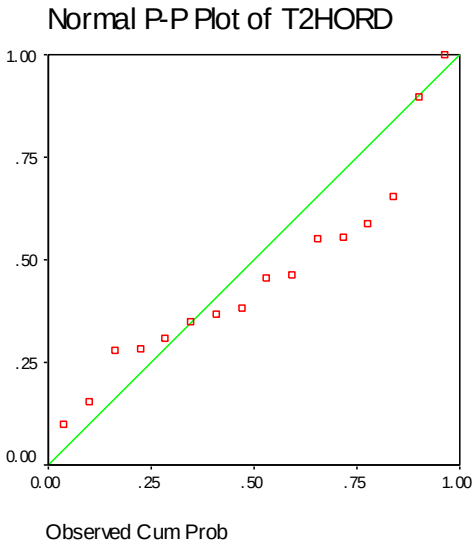
Key: T2REC – T2 at recall level

(v) Item cognitive ratings of T2 at comprehension level



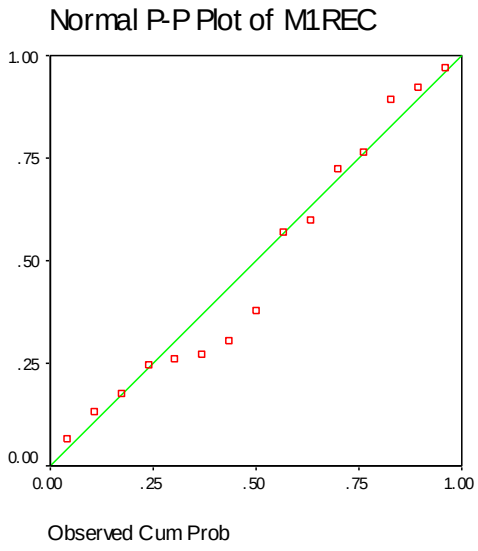
Key: T2COMP – T2 at comprehension level

(vi) Item cognitive ratings of T2 at higher order level



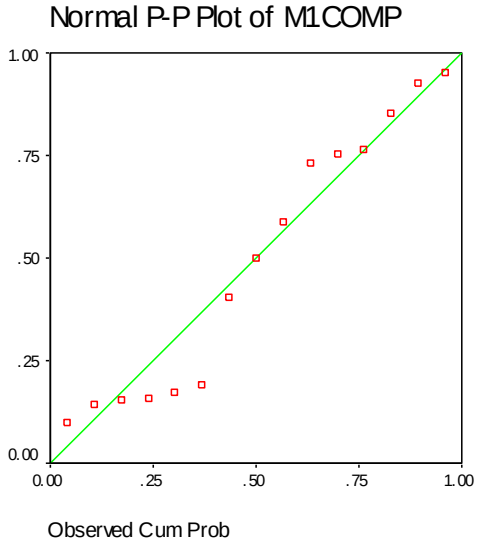
Key: T2HORD – T1 at higher order level

(vii) Item cognitive ratings of M1 at recall level



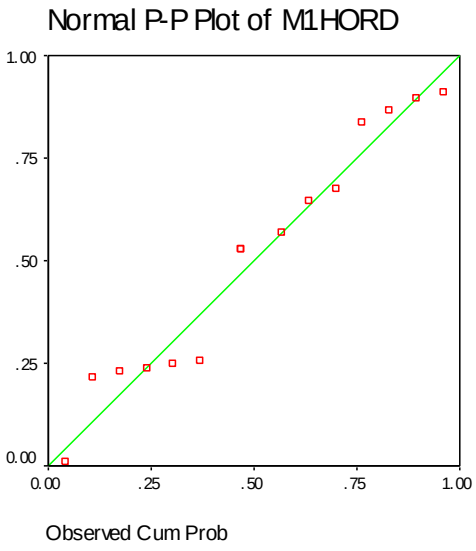
Key: M1REC – M1 at recall level

(viii) Item cognitive ratings of M1 at comprehension level



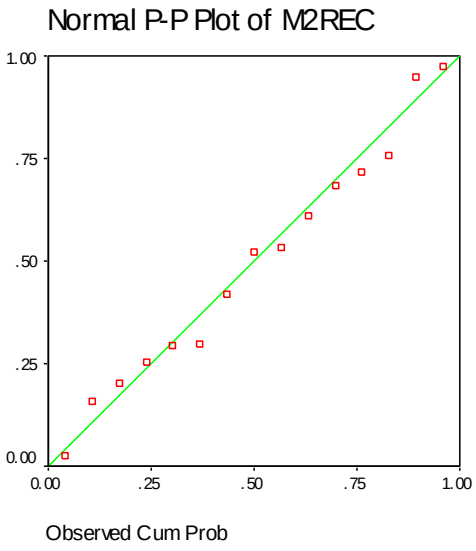
Key: M1COMP – M1 at comprehension level

(ix) Item cognitive ratings of M1 at higher order level

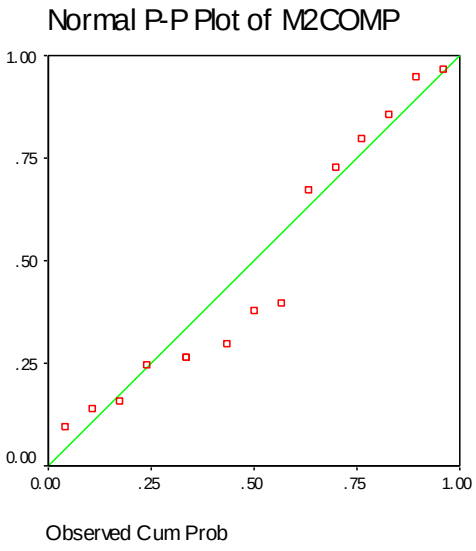


Key: M2HORD –M2 at higher order level

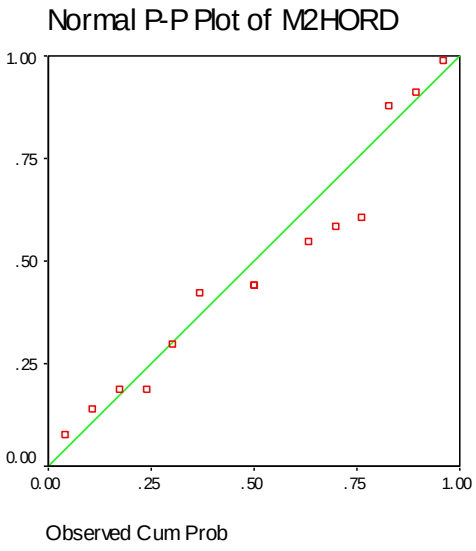
(x) Item cognitive ratings of M2 at recall level



(xi) Item cognitive ratings of M2 at comprehension level

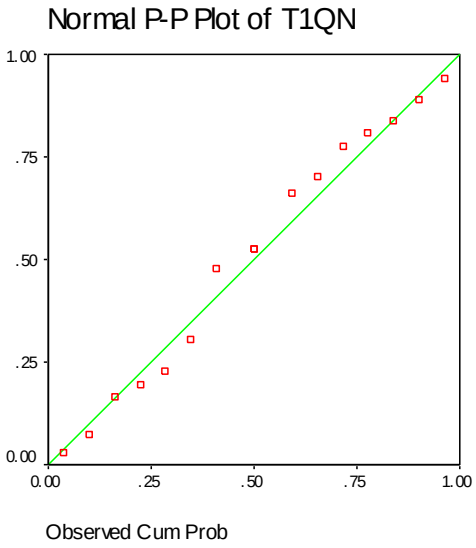


(xii) Item cognitive ratings of M2 at higher order level

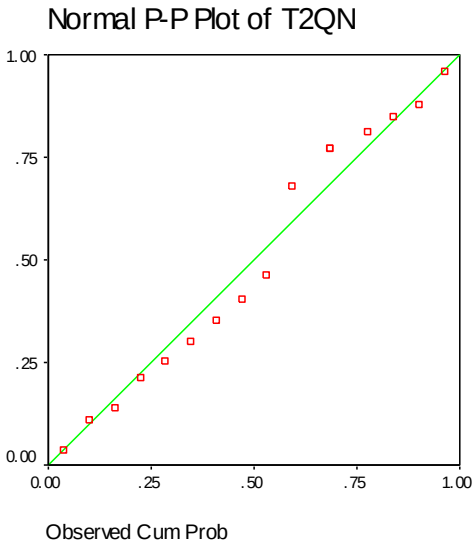


g. Exploratory factor analysis results of teachers' tests

(i) Total percentage variance explained for T1

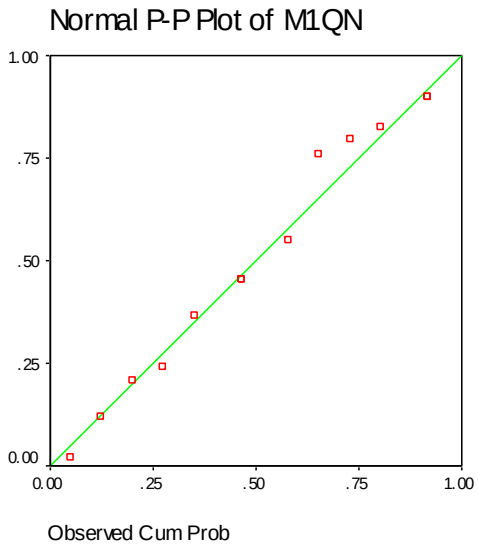


(ii) Total percentage variance explained for T2

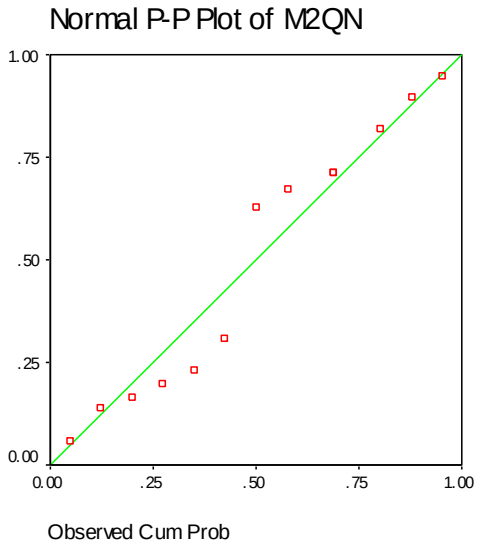


Key: T2QN – T2 at question level

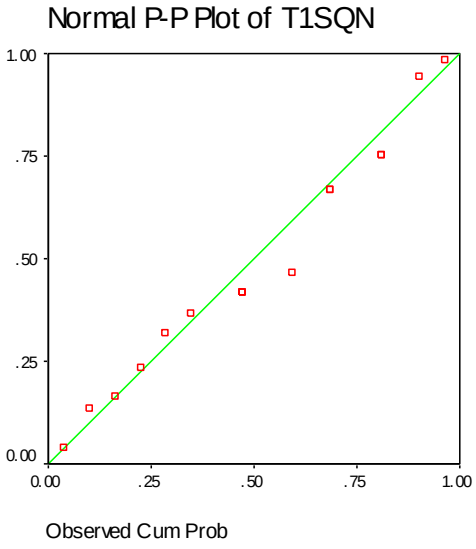
(iii) Total percentage variance explained for M1



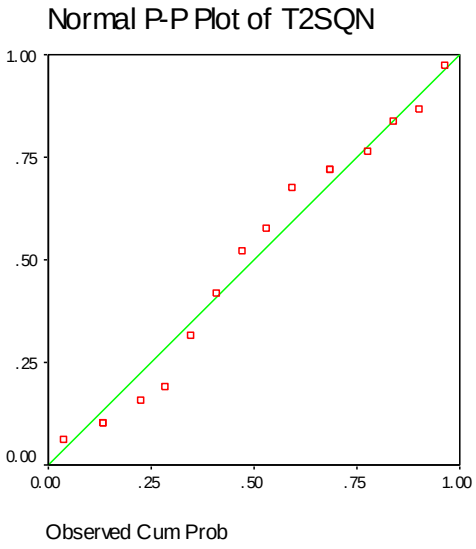
(iv) Total percentage variance explained for M2



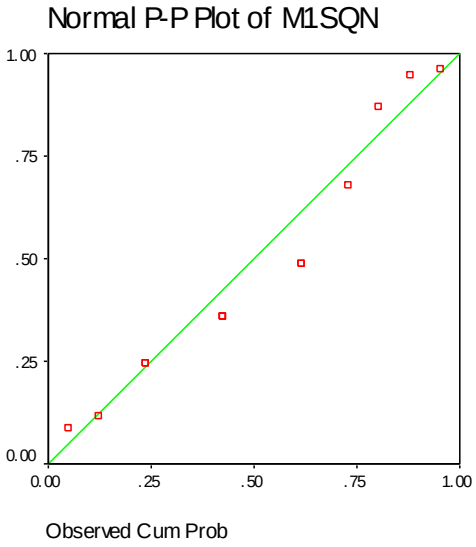
(v) Total percentage variance explained for T1



(vi) Total percentage variance explained for T2

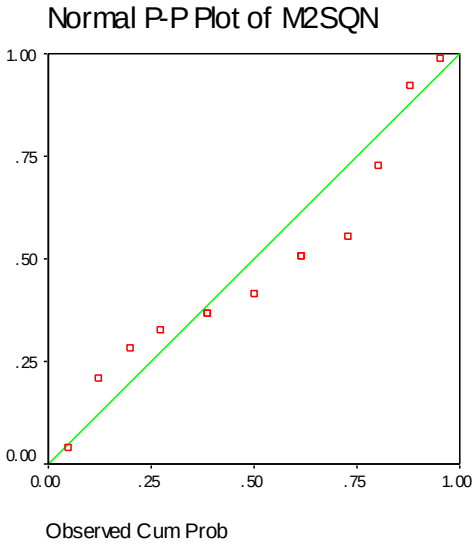


(vii) Total percentage variance explained for M1



Key: M1SQN – M1 at sub-question level

(viii) Total percentage variance explained for M2



APPENDIX 4.2

SIGNIFICANCE TEST RESULTS FOR QUALITY OF ITEMS AND RELIABILITY

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
					Lower	Upper			
Pair 1	T1gooditems - T2gooditems	.063	10.376	2.594	-5.467	5.592	.024	15	.981
Pair 2	M1gooditems - M2gooditems	-4.769	10.545	2.925	-11.141	1.603	-1.631	12	.129
Pair 3	T1alpha - T2alpha	.0007875	.0656539	.0164135	-.0341970	.0357720	.048	15	.962
Pair 4	M1alpha - M2alpha	.0122692	.0780576	.0216493	-.0349005	.0594390	.567	12	.581

APPENDIX 4.3

SIGNIFICANCE TEST RESULTS FOR DIFFICULTY OF ITEMS

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
					Lower	Upper			
Pair 1	T1difficult - T2difficult	-.688	24.069	6.017	-13.513	12.138	-.114	15	.911
Pair 2	T1acceptable - T2acceptable	.875	21.654	5.414	-10.664	12.414	.162	15	.874
Pair 3	T1easy - T2easy	-.188	6.025	1.506	-3.398	3.023	-.124	15	.903
Pair 4	M1difficult - M2difficult	-3.692	16.765	4.650	-13.823	6.439	-.794	12	.443
Pair 5	M1acceptable - M2acceptable	1.154	15.361	4.261	-8.129	10.437	.271	12	.791
Pair 6	M1easy - M2easy	2.538	3.307	.917	.540	4.537	2.768	12	.017

APPENDIX 4.4

SIGNIFICANCE TEST RESULTS ON USE OF PAST EXAMINATION ITEMS

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
					Lower	Upper			
Pair 1	T1pastquestions - T2pastquestions	11.375	25.492	6.373	-2.209	24.959	1.785	15	.095
Pair 2	M1pastquestions - M2pastquestions	7.143	23.274	6.220	-6.295	20.581	1.148	13	.272

APPENDIX 4.5

ITEM RELEVANCE RATING OF TEACHERS' PRE-TESTS AND POST-TESTS

Item review sheet: Teacher1/T1									
Mean item relevance ratings by SMEs									
	Topics						Cognitive levels		
Item Number	Properties of matter	Elements and chemical bonding	Chemical Reaction	Force and Motion	Organic chemistry	Electricity and magnetism	Recall	Comprehension	Higher Order
1a	1.50	4.00	2.00	1.00	1.00	1.17	3.17	2.67	1.00
1bi	1.33	4.33	2.67	1.00	1.00	1.00	1.83	3.00	2.17
1bii	1.33	3.67	2.33	1.00	1.00	1.00	1.67	3.00	1.83
1c	1.50	3.17	2.33	1.00	1.33	1.00	1.17	2.33	3.00
1d	1.00	4.33	2.00	1.00	1.17	1.00	3.00	2.50	1.33
1e	1.33	4.00	1.67	1.00	1.00	1.00	1.83	3.33	1.67
2ai	1.00	4.00	2.33	1.00	1.00	1.00	2.50	3.83	2.00
2aii	1.00	3.50	2.17	1.00	1.00	1.17	2.50	2.33	1.17
2bi	1.00	3.83	1.83	1.00	1.00	1.17	1.83	3.50	1.50
2bii	1.00	3.83	2.33	1.00	1.00	1.00	2.50	2.50	1.17
2biii	1.00	4.00	2.50	1.00	1.00	1.00	2.00	2.67	1.33
2biv	1.00	3.67	1.67	1.00	1.00	1.00	1.50	2.50	2.17
3ai	4.50	1.00	1.00	1.67	1.00	1.00	2.00	3.00	2.67
3aii	3.83	1.00	1.00	1.50	1.00	1.00	2.00	3.50	1.50
3aiii	4.17	1.00	1.00	1.33	1.00	1.00	3.00	2.00	1.50
3bi	4.00	1.17	1.00	1.17	1.00	1.00	3.33	1.67	1.00
3bii	4.17	1.00	1.00	1.33	1.00	1.00	1.50	2.67	2.17
3biii	4.33	1.00	1.00	1.33	1.00	1.00	1.17	2.67	3.17
3biv	3.50	1.00	1.00	1.17	1.00	1.00	1.67	2.33	1.67
3c	4.00	1.00	1.00	1.67	1.00	1.00	1.50	3.17	2.17

Item review sheet: Teacher1/T1									
Mean item relevance ratings by SMEs									
	Topics						Cognitive levels		
Item Number	Properties of matter	Elements and chemical bonding	Chemical Reaction	Force and Motion	Organic chemistry	Electricity and magnetism	Recall	Compre hension	Higher Order
4ai	1.33	1.00	1.17	4.67	1.00	1.00	1.67	3.33	1.17
3aii	1.33	1.00	1.17	4.17	1.00	1.00	1.00	1.83	3.83
4aiii	1.33	1.00	1.17	4.17	1.00	1.00	1.17	1.67	3.83
4bi	1.67	1.00	1.17	3.83	1.00	1.00	1.50	3.00	2.50
4bii	1.67	1.00	1.17	3.83	1.00	1.00	1.33	2.67	2.83
4biii	1.33	1.00	1.17	3.83	1.00	1.00	3.17	1.67	1.50
4biv	1.33	1.00	1.17	3.83	1.00	1.00	3.00	2.17	1.33
5i	1.33	3.83	2.17	1.17	1.33	1.00	1.50	2.67	2.33
5ii	1.17	2.17	2.00	1.00	3.67	1.00	2.00	2.83	1.67
6	1.20	1.40	4.20	1.00	1.20	1.00	1.25	2.75	3.25

Item review sheet: Teacher 1/T2									
Mean item relevance ratings by SMEs									
	Topics						Cognitive level		
Item Number	Properties of matter	Elements and chemical bonding	Chemical Reaction	Force and Motion	Organic chemistry	Electricity and magnetism	Recall	Compre hension	Higher Order
1ai	4.17	1.00	1.00	1.33	1.00	1.00	4.17	1.17	1.00
1aii	3.83	1.17	1.00	1.33	1.00	1.00	4.17	1.33	1.00
1bi	4.00	1.33	1.00	1.33	1.00	1.00	2.17	2.83	2.00
1bii	3.83	1.33	1.00	1.17	1.00	1.00	2.17	2.83	1.83
1c	4.17	1.00	1.00	1.17	1.17	1.00	1.17	2.17	4.00
2ai	1.50	3.17	1.33	1.00	1.00	1.33	3.50	1.17	1.00
2aii	1.50	3.83	1.33	1.00	1.00	1.00	1.50	2.67	2.17
2aiii	1.33	3.67	1.83	1.00	1.00	1.00	1.83	3.00	1.50
2aiv1	1.00	3.67	1.83	1.00	1.00	1.00	2.33	2.17	1.00
2aiv2	1.00	4.00	1.83	1.17	1.00	1.00	1.50	2.83	2.33
2aiv3	1.17	3.67	2.00	1.00	1.00	1.00	1.83	2.17	2.00
2av	1.17	3.17	2.67	1.00	1.00	1.00	1.83	2.83	2.17
2bi	1.17	3.67	2.17	1.00	1.00	1.17	2.17	2.67	2.17
2bii	1.67	2.83	1.67	1.00	1.00	1.50	1.83	2.67	1.83
2biii	1.83	2.17	2.17	1.00	1.00	1.17	2.50	2.67	1.33
2biv	2.00	2.33	1.33	1.00	1.00	1.00	3.50	1.33	1.00
2bv	2.33	3.17	1.33	1.17	1.00	1.00	1.50	3.00	1.83
2c	1.33	3.33	1.17	1.00	1.17	1.00	3.83	1.33	1.00
2di	1.50	3.17	1.83	1.00	1.17	1.00	2.00	2.50	2.17
2dii	1.50	3.17	1.50	1.00	1.00	1.00	4.00	1.33	1.00
2diii	1.50	3.17	1.17	1.00	1.00	1.00	1.83	2.50	2.17
2div	1.67	3.17	1.00	1.00	1.00	1.00	4.00	1.50	1.00
2dv	1.67	3.00	1.00	1.00	1.00	1.00	2.83	2.33	1.50
3ai	1.00	1.83	4.17	1.00	1.50	1.00	4.17	1.33	1.00
3aii	1.00	1.50	3.67	1.00	1.50	1.00	3.00	2.67	1.00
3aiii	1.17	1.17	4.17	1.00	1.33	1.00	3.50	1.33	1.50

Item review sheet: Teacher 1/T2									
Mean item relevance ratings by SMEs									
Item Number	Topics						Cognitive level		
	Properties of matter	Elements and chemical bonding	Chemical Reaction	Force and Motion	Organic chemistry	Electricity and magnetism	Recall	Compre hension	Higher Order
3aiv	1.17	1.50	4.17	1.00	1.33	1.00	4.00	1.17	1.00
3bi	1.17	1.67	4.33	1.00	1.50	1.00	1.17	2.33	3.17
3bii	1.17	1.33	4.00	1.00	1.33	1.00	1.33	2.33	3.50
3ci	1.17	1.00	3.67	1.00	1.50	1.00	4.00	1.33	1.00
3cii1	1.17	1.00	4.50	1.17	1.33	1.00	1.17	2.00	4.00
3cii2	1.17	1.00	4.50	1.17	1.33	1.00	1.33	1.50	4.00
3d	1.17	1.17	4.50	1.00	1.17	1.00	1.33	2.33	3.83
3e	2.00	1.00	3.17	1.00	1.33	1.00	3.17	1.33	1.00
3fi	1.83	1.00	3.33	1.00	1.33	1.00	1.17	2.17	4.00
3fii	1.83	1.17	3.17	1.00	1.33	1.00	1.17	2.17	3.83
3fiii	1.50	1.00	3.17	1.00	1.33	1.00	1.17	2.83	2.33
3gi	1.00	2.00	3.67	1.17	1.33	1.00	2.67	2.17	1.50
3gii	1.17	1.50	4.00	1.00	1.33	1.00	2.67	3.00	2.00
3giii	1.17	1.50	4.00	1.00	1.33	1.00	2.50	2.83	2.00
3hi	1.17	2.00	3.50	1.00	1.33	1.00	2.50	3.17	1.00
3hii	1.33	2.00	3.50	1.00	1.33	1.00	2.67	3.33	1.00
4a	1.00	1.00	1.00	4.33	1.00	1.00	3.00	2.33	1.67
4bi	1.00	1.00	1.00	4.17	1.00	1.00	1.17	2.33	3.33
4bii	1.00	1.00	1.00	4.00	1.00	1.00	1.33	2.17	3.83
4biv1	1.00	1.00	1.00	3.83	1.00	1.00	2.83	2.17	1.50
4biv2	1.00	1.00	1.00	4.00	1.00	1.00	2.67	2.17	1.67
4biv3	1.00	1.00	1.83	3.00	1.00	1.00	2.67	2.17	1.50
4ci	1.00	1.00	1.00	3.83	1.00	1.00	1.50	2.33	2.33
4cii	1.17	1.00	1.00	4.00	1.00	1.00	1.67	1.67	3.33
4ciii	1.17	1.00	1.00	4.17	1.00	1.00	1.33	2.00	3.33
4d	1.17	1.00	1.00	4.33	1.00	1.00	4.17	1.17	1.00
4e	1.17	1.00	1.00	4.33	1.00	1.00	3.67	1.00	1.00
4fi	1.33	1.00	1.00	4.00	1.00	1.00	2.00	1.67	3.33
4fii	1.17	1.00	1.00	4.50	1.00	1.00	1.50	1.83	3.67
5ai	1.17	1.83	1.00	1.00	3.33	1.00	2.67	2.00	1.33
5aii	1.00	1.67	1.00	1.00	3.33	1.00	1.83	3.00	1.50
5aiii	1.00	1.50	1.00	1.00	3.67	1.00	2.33	2.33	1.33
5b	1.00	1.33	2.17	1.00	2.67	1.00	1.67	2.50	2.50
5c	1.00	1.50	2.33	1.00	3.33	1.00	3.50	1.33	1.00
5di	1.00	1.50	1.17	1.00	3.50	1.00	2.17	3.00	1.17
5dii	1.00	1.67	1.00	1.00	3.50	1.00	1.50	3.33	1.67
5e	1.00	1.17	1.00	1.00	3.50	1.00	3.50	1.33	1.00
5fi	1.00	1.33	2.67	1.00	4.00	1.00	3.00	2.17	1.17
5fii1	1.00	1.17	2.50	1.00	3.67	1.00	2.50	2.67	1.33
5fii2	1.00	1.33	2.83	1.00	3.67	1.00	2.17	2.83	1.83
5fii3	1.00	1.17	2.17	1.00	4.00	1.00	3.50	2.17	1.00
5fiii	1.17	1.17	2.17	1.00	4.00	1.00	2.00	2.83	2.50
6a	1.00	1.50	1.00	1.00	4.17	1.00	4.00	1.33	1.00
6bi	1.00	1.50	1.00	1.00	4.00	1.00	2.00	3.17	1.50
6bii	1.00	1.50	1.00	1.00	3.67	1.00	1.83	3.17	1.33

Item review sheet: Teacher 1/T2									
Mean item relevance ratings by SMEs									
Item Number	Topics						Cognitive level		
	Properties of matter	Elements and chemical bonding	Chemical Reaction	Force and Motion	Organic chemistry	Electricity and magnetism	Recall	Compre hension	Higher Order
6ci	1.00	1.83	1.00	1.00	3.50	1.00	1.67	3.00	1.83
6cii	1.00	1.50	1.17	1.17	3.33	1.00	2.33	2.00	1.50
6ciii	1.00	1.33	1.17	1.00	3.50	1.00	1.50	3.00	1.50
6di	1.00	1.00	2.20	1.00	3.80	1.00	2.20	2.20	1.40
6dii1	1.00	1.17	2.17	1.00	3.67	1.00	2.67	2.50	1.17
6dii2	1.00	1.00	1.83	1.00	3.00	1.00	1.67	3.33	1.17
6dii3	1.00	1.00	2.00	1.00	2.67	1.00	2.83	1.67	1.00
6e	1.00	1.17	2.17	1.00	3.00	1.00	2.17	3.00	1.00
6f	1.00	1.00	1.83	1.00	3.17	1.00	1.50	3.17	1.50
6g	1.00	1.17	2.17	1.00	2.83	1.00	2.00	2.50	2.17
6hi	1.33	1.33	1.67	1.00	3.00	1.00	2.17	3.00	1.17
6hii	1.00	1.00	1.50	1.00	3.00	1.17	1.83	2.67	1.50
6i	1.33	1.17	1.67	1.00	3.17	1.00	2.67	2.50	1.33
7(1)	1.17	1.00	1.00	1.00	1.00	4.17	2.00	3.00	2.33
7(2)	1.17	1.00	1.00	1.00	1.00	4.17	2.00	3.17	2.17
7(3)	1.17	1.00	1.00	1.00	1.00	4.00	1.83	3.00	2.33
7i	1.00	1.00	1.00	1.17	1.00	3.83	1.33	2.50	3.00
7ii	1.00	1.00	1.00	1.17	1.00	4.00	1.17	1.67	3.50
7iii	1.00	1.00	1.00	1.17	1.00	3.83	1.33	1.67	3.50
7iv	1.00	1.00	1.00	1.17	1.00	3.83	1.33	1.67	3.50
7vi	1.00	1.00	1.00	1.17	1.00	3.83	1.33	3.67	3.50
7vi	1.00	1.00	1.00	1.17	1.00	3.83	1.17	2.00	3.17
8	3.83	1.50	1.00	1.33	1.00	1.00	1.67	2.33	2.50
10	1.33	1.00	1.33	1.00	2.67	1.50	1.83	2.33	2.67

tem review sheet: Teacher 2/T1							
Mean item relevance ratings by SMEs							
	Topics				Cognitive level		
	Properties of matter	Chemical Reaction	Force and Motion	Organic Chemistry	Recall	Comprehension	Higher Order
1ai	1.40	1.60	1.00	1.00	2.60	3.00	1.00
1aii	1.20	1.40	1.00	1.00	2.40	3.20	1.60
1aiii	1.20	2.00	1.00	1.00	2.40	3.20	1.40
1aiv	1.00	1.60	1.00	1.00	2.20	3.20	1.40
1av	1.20	2.00	1.00	1.00	2.40	3.20	1.20
1avi	1.00	1.80	1.00	1.00	2.20	2.80	1.60
1avii	1.20	1.40	1.00	1.00	1.80	3.60	1.80
1bi	1.60	1.40	1.00	1.00	3.20	2.00	1.00
1bii	1.60	1.20	1.00	1.00	3.00	2.20	1.00
1biii	2.20	1.40	1.20	1.00	2.00	2.80	1.80
1biv	2.20	1.40	1.00	1.00	2.20	2.60	1.80
Ci	1.00	2.60	1.00	1.00	2.40	2.40	1.60
Cii	1.20	2.00	1.00	1.00	2.20	3.40	2.40
D	1.40	1.40	1.00	1.00	2.80	2.60	1.60
2a	1.00	1.00	4.20	1.00	4.00	1.60	1.00

Item review sheet: Teacher 1/T2									
Mean item relevance ratings by SMEs									
Item Number	Topics						Cognitive level		
	Properties of matter	Elements and chemical bonding	Chemical Reaction	Force and Motion	Organic chemistry	Electricity and magnetism	Recall	Compre hension	Higher Order
2b	1.00	1.00	3.80	1.00	4.20	1.60	1.00		
2ci	1.20	1.00	4.20	1.00	3.00	2.80	1.20		
2cii	1.20	1.00	3.60	1.00	2.80	2.20	1.00		
2ciii	1.00	1.00	3.80	1.00	2.40	3.80	1.40		
2d	1.20	1.00	3.80	1.00	3.20	2.00	2.00		
2ei	1.00	1.00	4.40	1.00	1.80	3.60	2.20		
2eii	1.00	1.00	4.60	1.00	1.40	2.60	4.20		
2eiii	1.00	1.00	4.40	1.00	1.40	2.60	4.20		
2fi	1.20	1.00	4.40	1.00	1.60	3.20	3.20		
2fii	1.20	1.00	4.00	1.00	1.60	3.40	2.80		
2fiii	1.20	1.00	3.40	1.00	2.60	2.40	1.60		
3ai	3.40	1.20	1.60	1.00	3.60	2.00	1.00		
3aii	3.40	1.00	1.60	1.00	2.00	2.80	1.40		
3bi	4.00	1.00	1.40	1.00	4.20	1.20	1.00		
3bii	4.20	1.00	1.20	1.00	4.00	1.80	1.00		
3ci	4.00	1.00	1.00	1.00	3.60	1.80	1.80		
3cii	4.60	1.00	1.00	1.00	2.20	1.60	3.00		
3d	4.20	1.00	1.20	1.00	3.80	1.60	1.20		
4ai	1.00	1.00	1.00	3.60	4.00	1.20	1.00		
4aii	1.00	1.00	1.00	3.40	3.80	1.60	1.00		
4bi	1.20	1.00	1.00	3.80	3.60	2.20	1.00		
4bii	1.00	2.20	1.00	3.40	3.20	2.40	1.00		
4biii	1.00	2.20	1.00	3.40	2.80	2.80	1.40		
4biv	1.00	1.20	1.00	3.80	1.60	3.00	2.40		
4bv	1.00	1.40	1.00	3.20	2.20	2.80	1.40		
4bvi	1.40	1.40	1.00	3.60	2.20	2.80	1.60		
4bvii	1.00	1.20	1.00	3.60	1.80	2.80	1.60		
5ai	1.00	2.00	1.00	3.40	2.80	2.60	1.00		
5aii	1.00	2.60	1.00	2.80	1.60	2.60	2.60		
5aiii	1.00	2.60	1.00	3.00	1.80	3.00	2.00		
5b	1.40	3.60	1.00	1.60	1.20	2.20	3.80		
5ci	1.00	3.80	1.00	1.40	3.80	1.20	1.00		
5cii	1.00	4.20	1.00	2.00	1.20	2.00	4.20		
5di	1.20	3.80	1.00	1.20	4.20	1.00	1.20		
5dii	1.00	4.20	1.00	1.20	1.40	2.20	4.20		
6a	1.00	3.40	1.00	1.20	1.60	3.20	3.00		
6b	1.20	4.00	1.00	1.20	1.80	2.40	3.60		
7	1.20	1.00	4.40	1.00	1.20	2.20	4.20		

Item review sheet: Teacher 2/T2							
Mean item relevance ratings by SMEs							
	Topics				Cognitive level		
Item number	Properties of matter	Chemical Reaction	Force and Motion	Organic Chemistry	Recall	Compreh.	Higher Order
1a	1.00	1.00	3.67	1.00	3.67	2.00	1.00
1b	1.50	1.00	3.83	1.00	2.33	3.33	1.83
1c	1.00	1.00	4.00	1.00	4.33	1.33	1.00
1di	1.00	1.00	3.67	1.00	3.83	1.17	1.17
1dii	1.00	1.00	3.50	1.00	2.33	2.67	1.67
1diii	1.00	1.00	3.67	1.17	3.67	2.00	1.00
1div	1.00	1.00	3.50	1.00	1.50	3.00	1.83
1e	1.00	1.00	4.33	1.00	1.50	2.17	4.17
1fi	1.17	1.00	4.17	1.00	1.33	1.50	4.17
1fii	1.00	1.00	4.17	1.00	1.33	1.33	4.17
2ai	3.83	1.17	1.17	1.00	1.33	3.00	2.17
2aii	3.67	1.17	1.17	1.00	1.67	3.33	1.67
2bi	4.33	1.00	1.17	1.00	4.33	1.00	1.17
2bii	4.67	1.00	1.00	1.00	1.33	1.67	4.50
2ci	3.50	1.00	1.67	1.00	1.17	1.83	3.67
2cii	3.50	1.00	1.50	1.00	1.50	2.83	2.50
2di	3.83	1.00	1.33	1.00	1.67	2.50	1.67
2dii	3.83	1.00	1.17	1.00	1.83	3.00	1.50
2diii	4.00	1.00	1.17	1.00	1.67	2.67	1.83
2div	4.00	1.00	1.17	1.00	1.83	3.17	1.67
2ei	4.33	1.00	1.17	1.00	1.33	2.67	3.33
2eii	3.83	1.00	1.33	1.00	1.33	3.50	1.33
3ai	1.00	3.67	1.00	1.17	1.33	3.00	2.83
3aii	1.00	4.17	1.00	1.17	1.17	2.83	3.67
3aiii	1.00	4.00	1.00	1.00	1.33	2.00	4.33
3bi	1.00	4.00	1.00	1.83	1.17	2.17	3.67
3bii	1.00	3.67	1.00	2.17	1.50	2.33	3.50
3ci	1.00	3.83	1.00	1.33	1.17	1.83	3.50
3cii	1.00	3.67	1.00	1.33	1.17	1.67	3.33
4ai	1.00	1.50	1.00	3.33	2.83	2.33	1.50
4aii	1.00	1.00	1.50	3.50	3.50	1.50	1.00
4aiii	1.00	1.00	1.00	4.17	1.83	3.00	2.17
4b	1.00	4.33	1.00	2.17	1.33	3.17	3.00

Note: More data on mean item relevance rating by SMEs for other tests is on the CD

APPENDIX 4.6

SIGNIFICANCE TEST RESULTS FOR ITEM RELEVANCE AND REPRESENTATIVENESS

Paired Samples Test

		Paired Differences					T	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
					Lower	Upper			
Pair 1	T1relevance - T2relevance	.189	1.902	.491	-.864	1.242	.386	14	.706
Pair 2	M1relevance - M2relevance	-4.581	9.882	2.552	-10.054	.891	-1.795	14	.094
Pair 3	M1relevance - M2relevance	-4.581	9.882	2.552	-10.054	.891	-1.795	14	.094
Pair 4	T1representative - T2representative	-1.500	2.338	.585	-2.746	-.254	-2.566	15	.021

APPENDIX 4.7

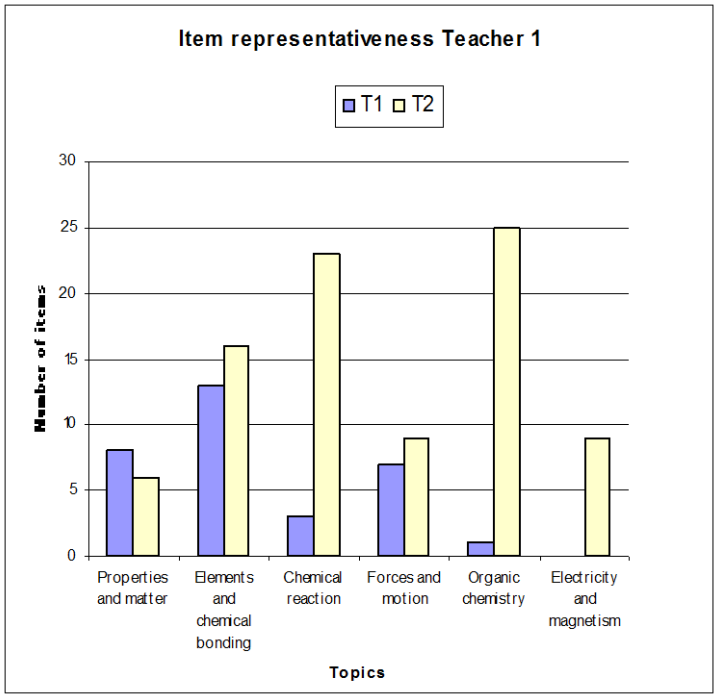
SIGNIFICANCE TEST RESULTS FOR COGNITIVE LEVEL OF ITEMS

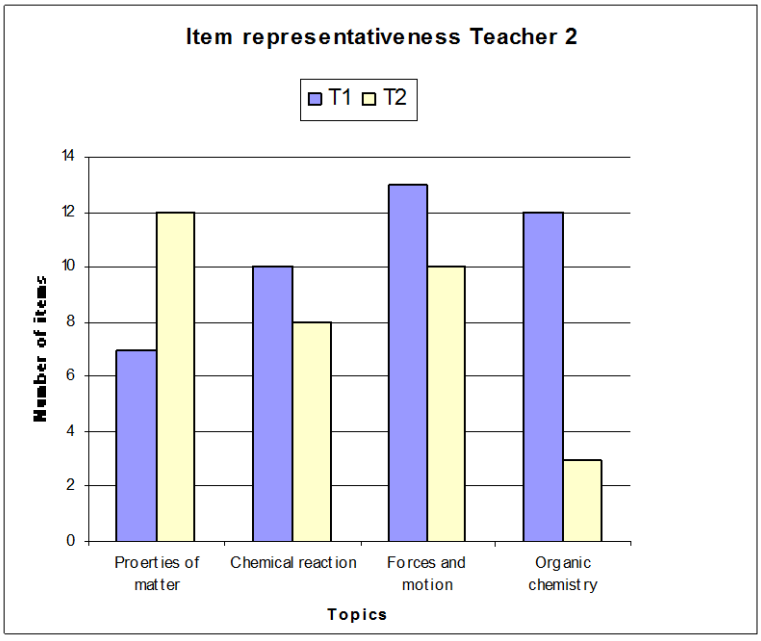
Paired Samples Test

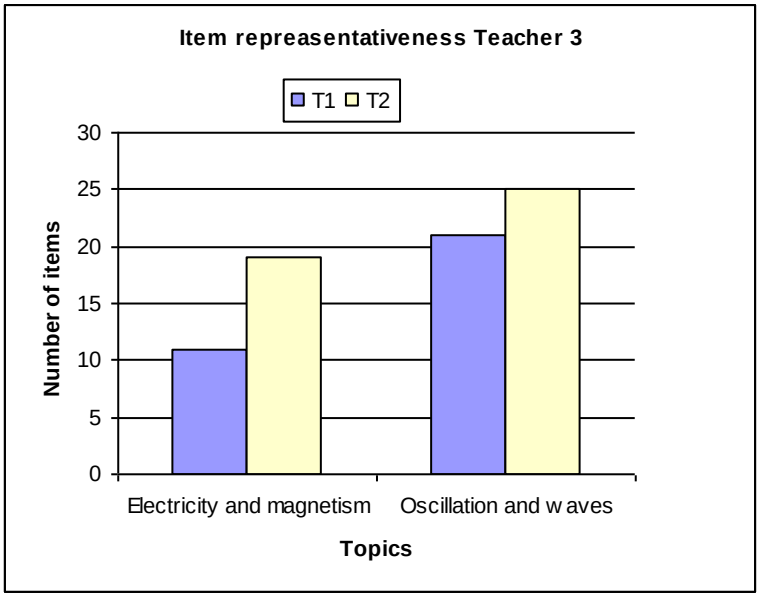
		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
					Lower	Upper			
Pair 1	T1recall - T1compreh	-4.34688	16.73805	4.18451	-13.26595	4.57220	-1.039	15	.315
Pair 2	T1recall - T1higherorder	7.03687	14.13786	3.53447	-.49666	14.57041	1.991	15	.065
Pair 3	T1comprehension - T1higherord	11.38375	15.26470	3.81618	3.24976	19.51774	2.983	15	.009
Pair 4	T2recall - T2compreh	-2.32625	15.07631	3.76908	-10.35985	5.70735	-.617	15	.546
Pair 5	T2recall - T2higherord	8.35312	15.14617	3.78654	.28230	16.42395	2.206	15	.043
Pair 6	T2compreh - T2higherord	10.67938	19.24902	4.81226	.42230	20.93645	2.219	15	.042
Pair 7	M1recall - M1compreh	.80000	23.94218	6.18184	-12.45874	14.05874	.129	14	.899
Pair 8	M1recall - M1higheror	15.79400	16.94508	4.37520	6.41013	25.17787	3.610	14	.003
Pair 9	M1comprehension - M1higheror	14.99400	12.68120	3.27427	7.97138	22.01662	4.579	14	.000
Pair 10	M2recall - M2compre	-3.57933	12.00312	3.09919	-10.22644	3.06777	-1.155	14	.267
Pair 11	M2recall - M2higheror	8.92867	15.49109	3.99978	.34999	17.50735	2.232	14	.042
Pair 12	M2compreh -M2higheror	12.50800	10.27909	2.65405	6.81563	18.20037	4.713	14	.000
Pair 13	T1recall - T2recall	-1.17063	10.08054	2.52013	-6.54217	4.20092	-.465	15	.649
Pair 14	T1compreh - T2compreh	.85000	10.67165	2.66791	-4.83652	6.53652	.319	15	.754
Pair 15	T1higherore - T2higherord	.14562	9.43536	2.35884	-4.88212	5.17337	.062	15	.952
Pair 16	M1recall - M2recall	3.74933	11.72048	3.02622	-2.74125	10.23992	1.239	14	.236
Pair 17	M1compre - M2compreh	-.63000	15.03257	3.88139	-8.95476	7.69476	-.162	14	.873
Pair 18	M1higheror - M2higheror	-3.11600	6.82997	1.76349	-6.89831	.66631	-1.767	14	.099

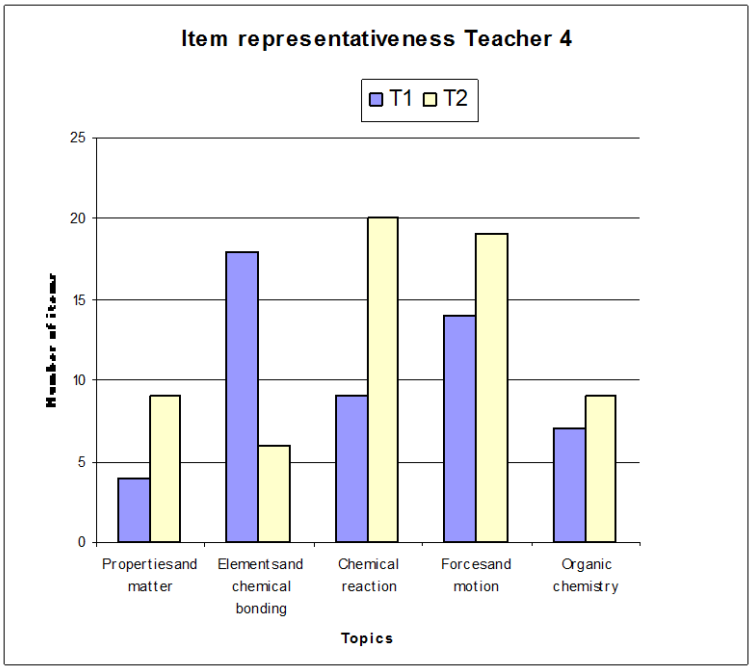
APPENDIX 4.8

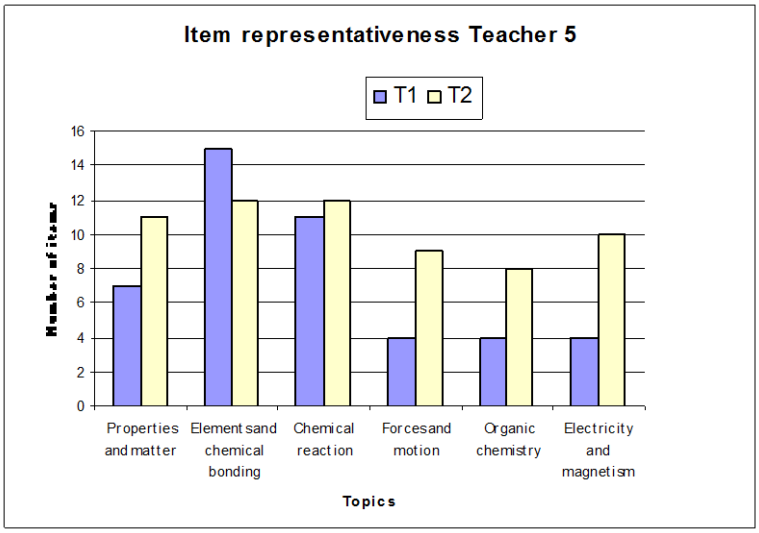
GRAPHICAL PRESENTATION OF ITEM REPRESENTATIVENESS AT TOPIC LEVEL

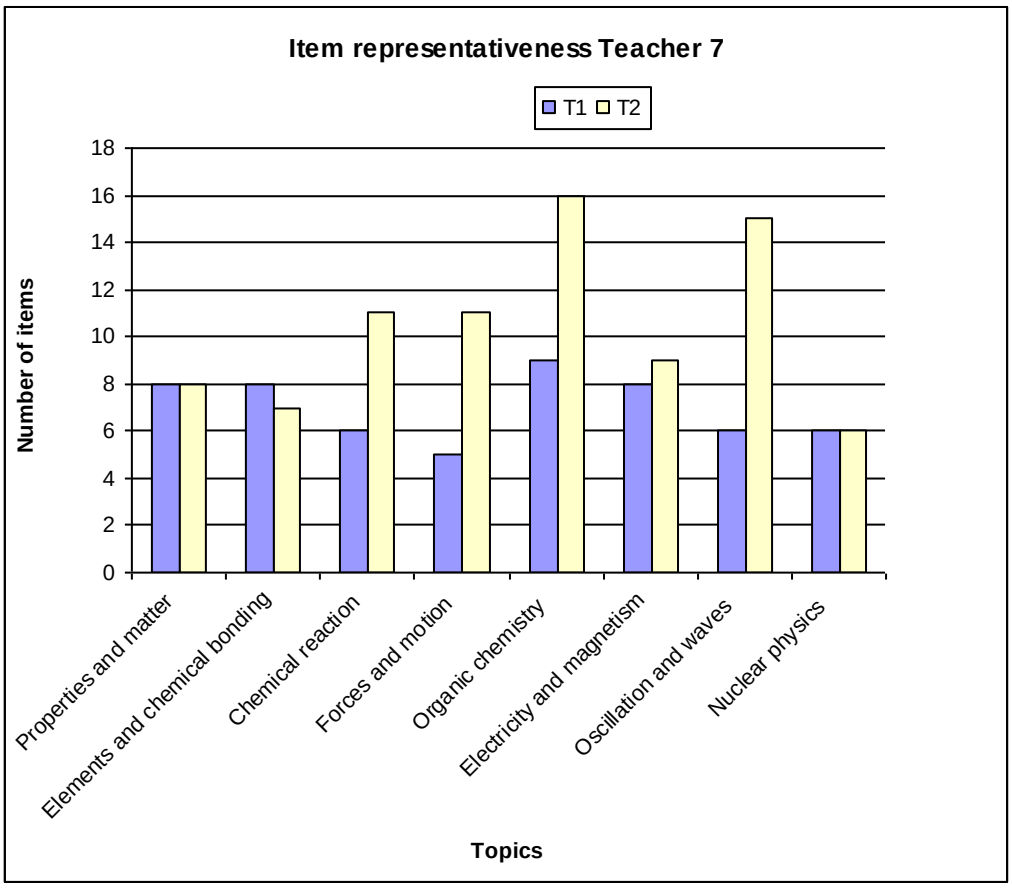


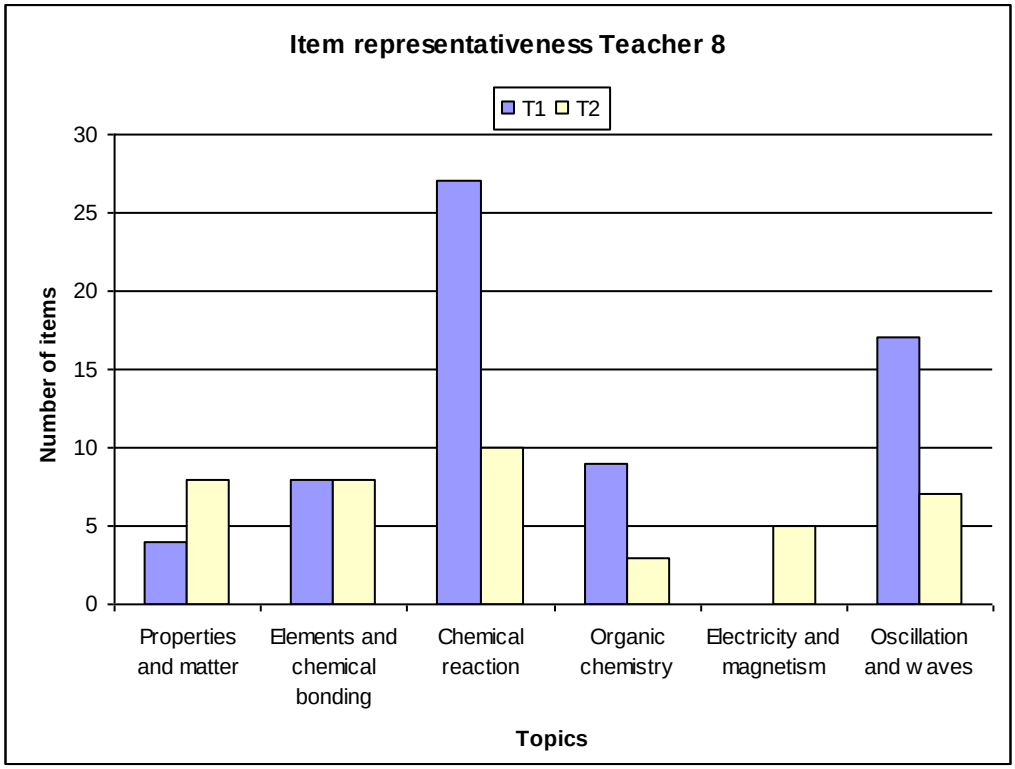


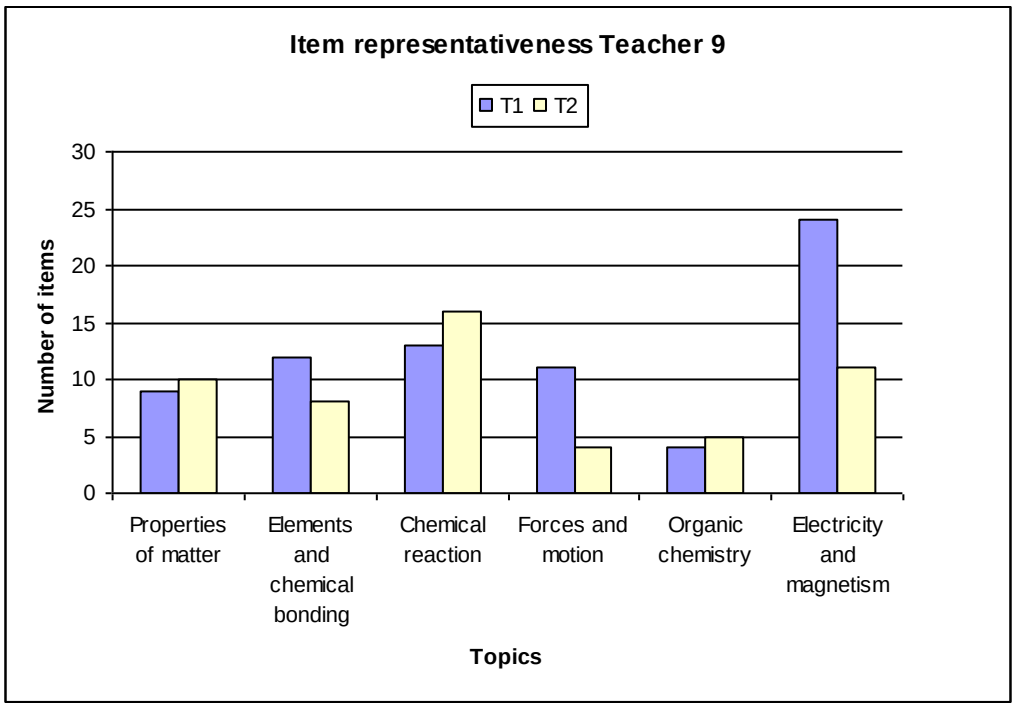

Item repreasentativeness Teacher 3
T1 T2
30
25
20
15
10
5
0
Number of items
Electricity and magnetism
Oscillation and w aves
Topics

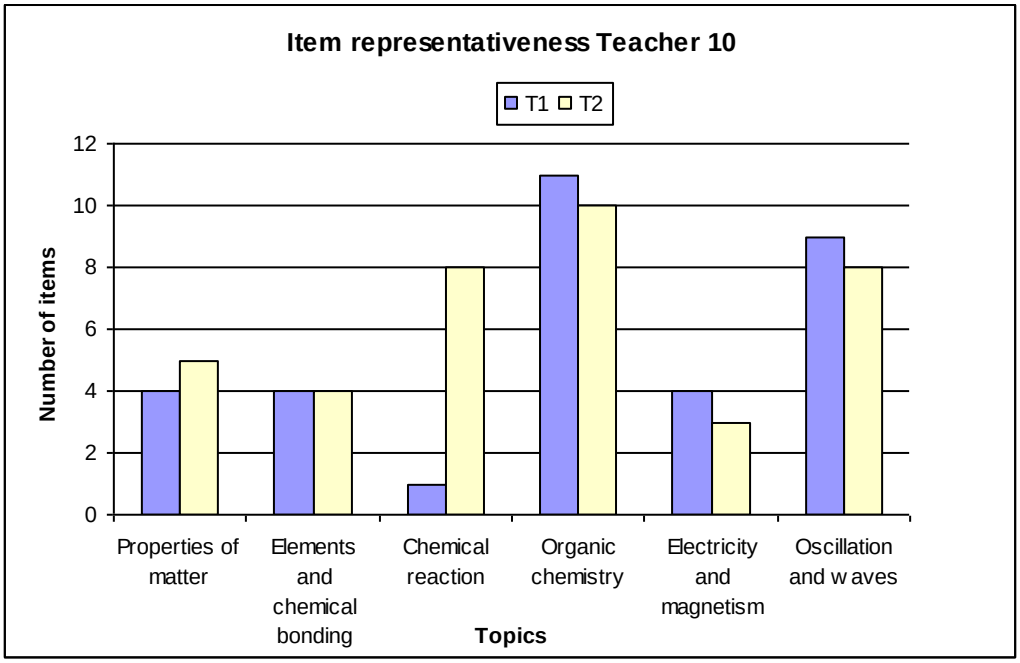

Item representativeness Teacher 4
T1 T2
25
20
15
10
5
0
Properties and matter
Elements and chemical bonding
Chemical reaction
Forces and motion
Organic chemistry
Topics

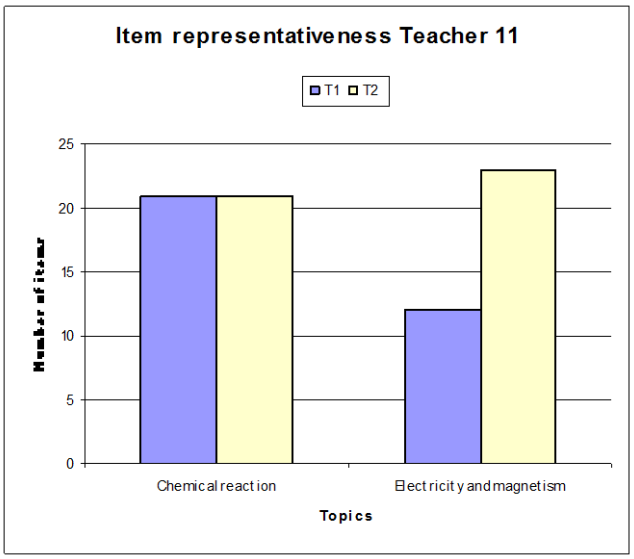

Item representativeness Teacher 5
T1
T2
16
14
12
10
8
6
4
2
0
Properties and matter
Elements and chemical bonding
Chemical reaction
Forces and motion
Organic chemistry
Electricity and magnetism
Topics

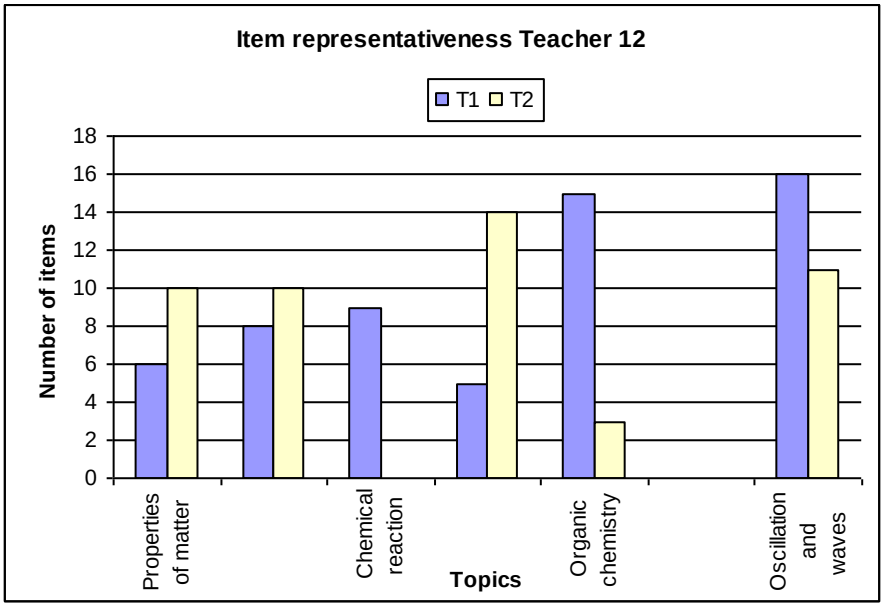

Item representativeness Teacher 7
T1
T2
Number of items
18
16
14
12
10
8
6
4
2
0
Properties and matter
Elements and chemical bonding
Chemical reaction
Forces and motion
Organic chemistry
Electricity and magnetism
Oscillation and waves
Nuclear physics
Topics

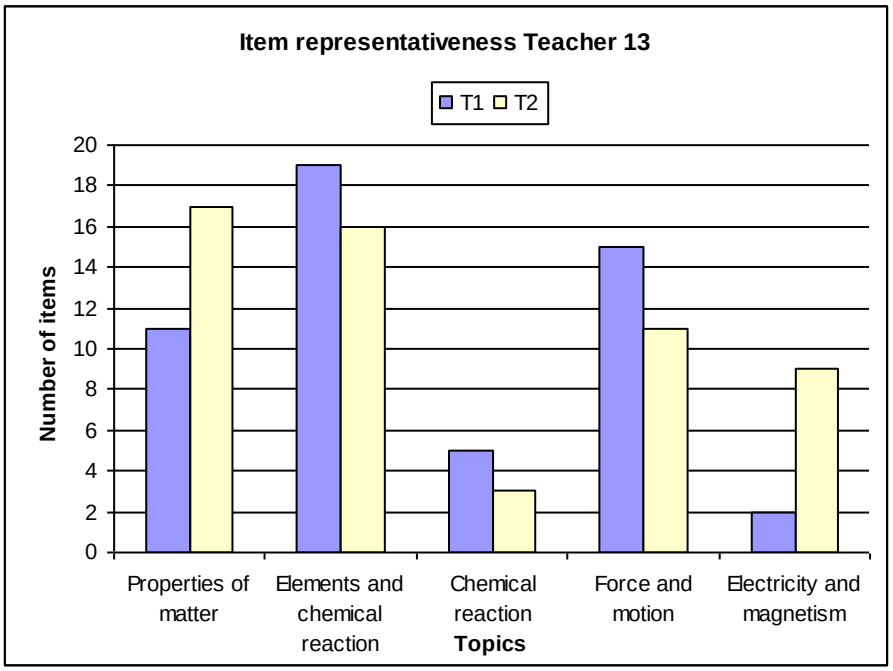

Item representativeness Teacher 8
T1
T2
30
25
20
15
10
5
0
Number of items
Properties and matter
Elements and chemical bonding
Chemical reaction
Organic chemistry
Electricity and magnetism
Oscillation and w aves
Topics

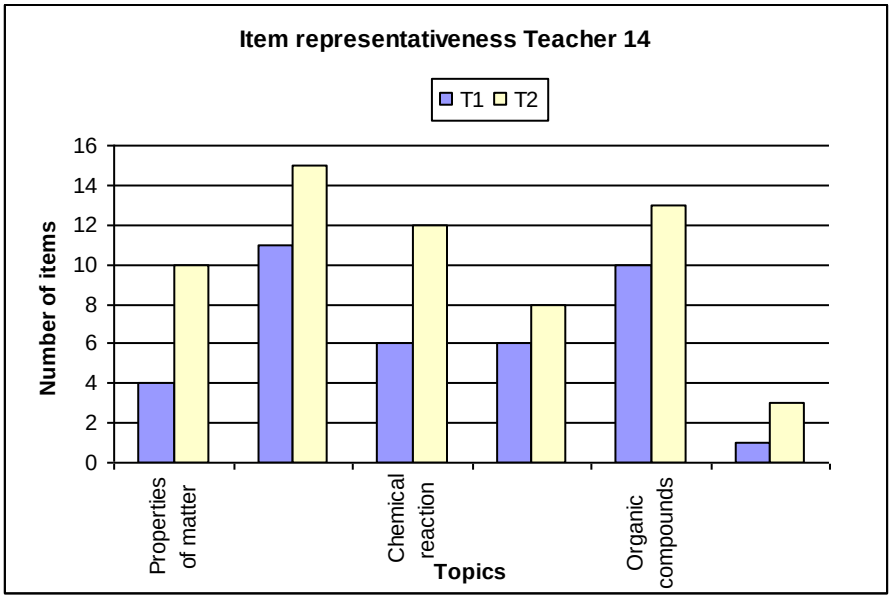

Item representativeness Teacher 9
T1
T2
30
25
20
15
10
5
0
Number of items
Properties of matter
Elements and chemical bonding
Chemical reaction
Forces and motion
Organic chemistry
Electricity and magnetism
Topics

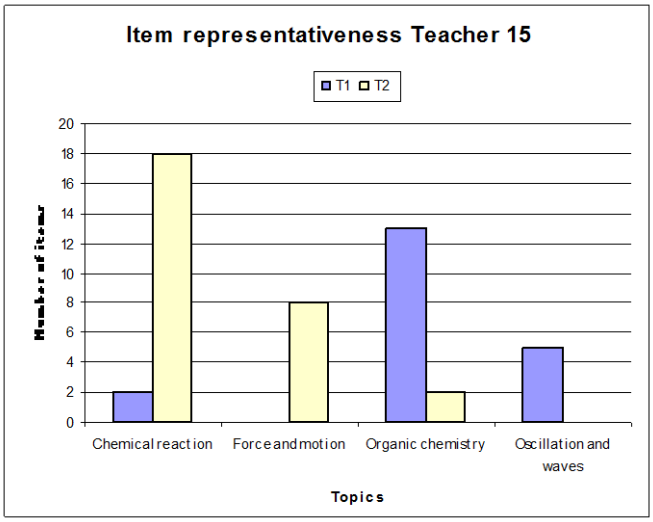

Item representativeness Teacher 10
T1
T2
12
10
8
6
4
2
0
Number of items
Properties of matter
Elements and chemical bonding
Chemical reaction
Organic chemistry
Electricity and magnetism
Oscillation and w aves
Topics

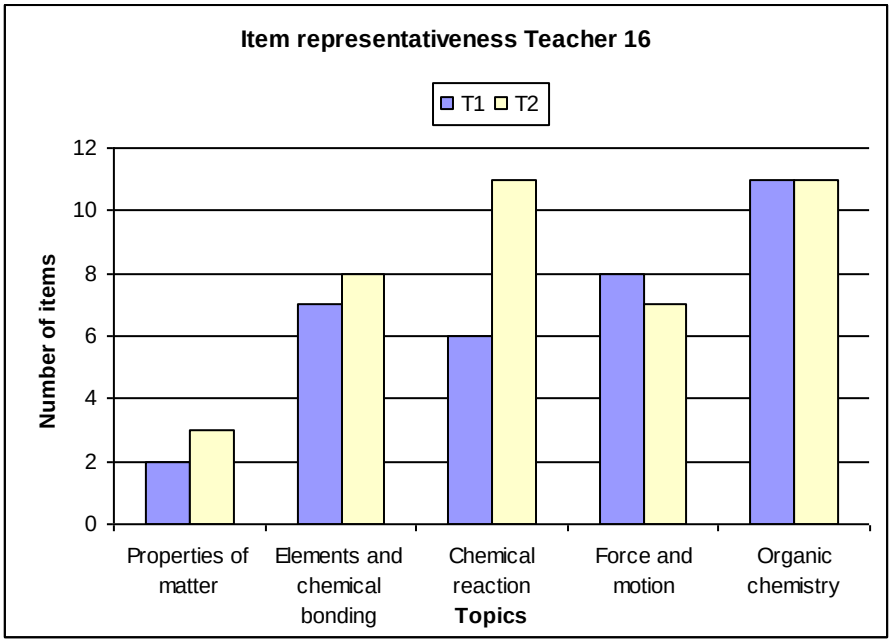

Item representativeness Teacher 11
T1
T2
25
20
15
10
5
0
Chemical reaction
Electricity and magnetism
Topics

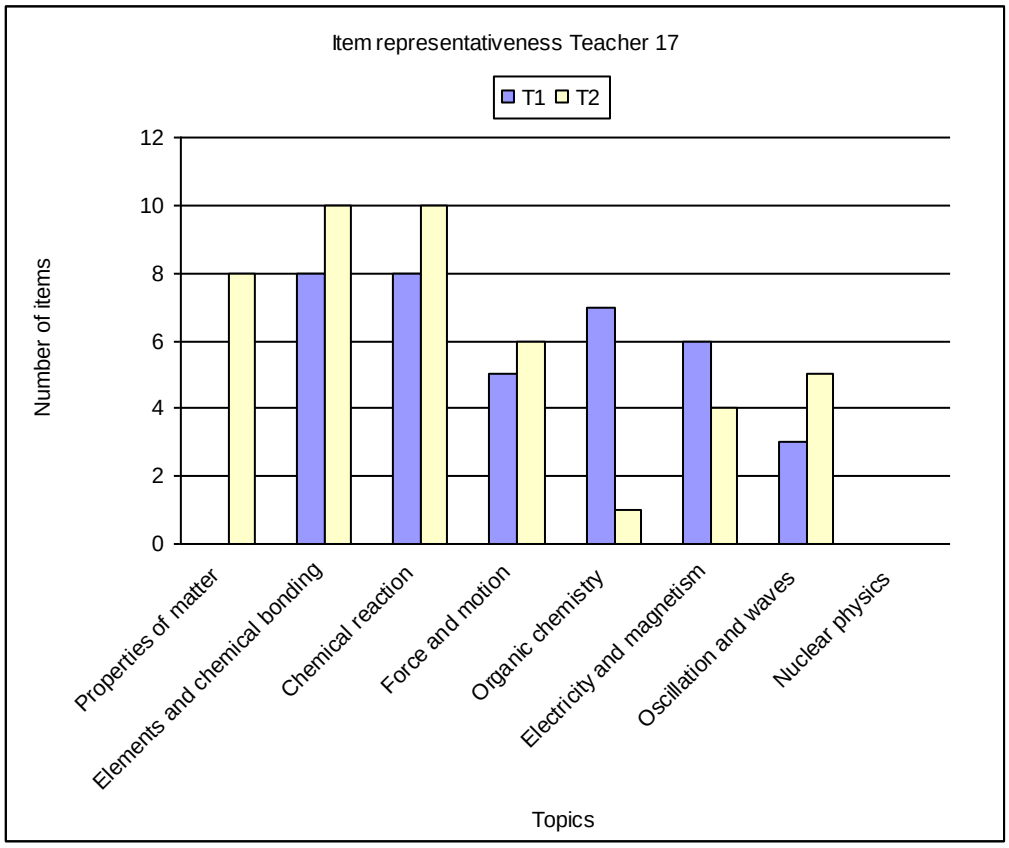

Item representativeness Teacher 12
T1
T2
Number of items
0
2
4
6
8
10
12
14
16
18
Properties of matter
Chemical reaction
Organic chemistry
Oscillation and waves
Topics


Item representativeness Teacher 13
T1
T2
Number of items
0
2
4
6
8
10
12
14
16
18
20
Properties of matter
Elements and chemical reaction
Chemical reaction
Force and motion
Electricity and magnetism
Topics


Item representativeness Teacher 14
T1 T2
Number of items
16
14
12
10
8
6
4
2
0
Properties of matter
Chemical reaction
Organic compounds
Topics


Item representativeness Teacher 15
T1 T2
20
18
16
14
12
10
8
6
4
2
0
Chemical reaction
Force and motion
Organic chemistry
Oscillation and waves
Topics


Item representativeness Teacher 16
T1
T2
Number of items
12
10
8
6
4
2
0
Properties of matter
Elements and chemical bonding
Chemical reaction
Force and motion
Organic chemistry
Topics


Item representativeness Teacher 17
T1
T2
Number of items
12
10
8
6
4
2
0
Properties of matter
Elements and chemical bonding
Chemical reaction
Force and motion
Organic chemistry
Electricity and magnetism
Oscillation and waves
Nuclear physics
Topics

APPENDIX 4.9

SAMPLE EFA RESULTS

a. Question level

Teacher 1 /test 1

Correlation Matrix(a)

		1	2	3	4	5	6
Sig. (1-tailed)	1		.000	.130	.033	.061	.104
	2	.000		.000	.000	.172	.010
	3	.130	.000		.007	.026	.117
	4	.033	.000	.007		.439	.076
	5	.061	.172	.026	.439		.044
	6	.104	.010	.117	.076	.044	

a Determinant = .450

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.645
Bartlett's Test of Sphericity	Approx. Chi-Square	69.661
	Df	15
	Sig.	.000

Communalities

	Initial	Extraction
1	.152	.144
2	.422	.893
3	.226	.223
4	.253	.297
5	.094	.999
6	.087	.094

Extraction Method: Principal Axis Factoring.

Total Variance Explained

Factor	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings(a)
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total
1	2.132	35.538	35.538	1.678	27.971	27.971	1.621
2	1.090	18.168	53.707	.972	16.208	44.179	1.078
3	.888	14.803	68.510				
4	.862	14.364	82.874				
5	.620	10.329	93.203				
6	.408	6.797	100.000				

Extraction Method: Principal Axis Factoring.
a When factors are correlated, sums of squared loadings cannot be added to obtain a total variance.

Factor Matrix(a)

	Factor	
	1	2
1		
2	.892	
3		
4		
5		.902
6		

Extraction Method: Principal Axis Factoring.

a Attempted to extract 2 factors. More than 102 iterations required. (Convergence=.003). Extraction was terminated.

Note that in 103 iterations extraction was not possible

Factor Matrix(a)

a Attempted to extract 2 factors. In iteration 103, the communality of a variable exceeded 1.0. Extraction was terminated.

Pattern Matrix(a)

	Factor	
	1	2
1		
2	.949	
3		
4	.547	
5		.993
6		

Extraction Method: Principal Axis Factoring. Rotation Method: Promax with Kaiser Normalization.
a Rotation converged in 3 iterations.

Structure Matrix

	Factor	
	1	2
1		
2	.944	
3		
4	.536	
5		.998
6		

Extraction Method: Principal Axis Factoring. Rotation Method: Promax with Kaiser Normalization.

Factor Correlation Matrix

Factor	1	2
1	1.000	.109
2	.109	1.000

Extraction Method: Principal Axis Factoring. Rotation Method: Promax with Kaiser Normalization.

Teacher 1/ test 2

Correlation Matrix(a)

		1	2	3	4	5	6	7	8	9
Sig. (1-tailed)	1		.000	.000	.000	.000	.000	.064	.002	.086
	2	.000		.000	.000	.000	.000	.003	.001	.003
	3	.000	.000		.000	.000	.000	.003	.000	.001
	4	.000	.000	.000		.000	.000	.000	.000	.000
	5	.000	.000	.000	.000		.000	.000	.029	.000
	6	.000	.000	.000	.000	.000		.000	.016	.000
	7	.064	.003	.003	.000	.000	.000		.043	.000
	8	.002	.001	.000	.000	.029	.016	.043		.001
	9	.086	.003	.001	.000	.000	.000	.000	.001	

a Determinant = .019

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.874
Bartlett's Test of Sphericity	Approx. Chi-Square	422.699
	Df	36
	Sig.	.000

Communalities

	Initial	Extraction
1	.457	.563
2	.591	.644
3	.630	.725
4	.568	.610
5	.589	.599
6	.552	.569
7	.181	.223
8	.233	.152
9	.343	.555

Extraction Method: Principal Axis Factoring.

Total Variance Explained

Factor	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings(a)
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total
1	4.406	48.955	48.955	3.988	44.314	44.314	3.770
2	1.138	12.643	61.598	.652	7.249	51.564	2.837
3	.934	10.377	71.975				
4	.709	7.882	79.857				
5	.458	5.089	84.946				
6	.408	4.530	89.476				
7	.376	4.179	93.655				
8	.325	3.609	97.263				
9	.246	2.737	100.000				

Extraction Method: Principal Axis Factoring.
a When factors are correlated, sums of squared loadings cannot be added to obtain a total variance.

Pattern Matrix(a)

	Factor	
	1	2
1	.878	
2	.790	
3	.849	
4	.670	
5		
6		.512
7		
8		
9		.854

Extraction Method: Principal Axis Factoring. Rotation Method: Promax with Kaiser Normalization.
a Rotation converged in 3 iterations.

Structure Matrix

	Factor	
	1	2
1	.728	
2	.802	.519
3	.851	.541
4	.771	.583
5	.718	.678
6	.642	.713
7		
8		
9		.729

Extraction Method: Principal Axis Factoring. Rotation Method: Promax with Kaiser Normalization.

Factor Correlation Matrix

Factor	1	2
1	1.000	.632
2	.632	1.000

Extraction Method: Principal Axis Factoring. Rotation Method: Promax with Kaiser Normalization.

Note: Results of EFA for the rest of the tests are summarized on the CD.

APPENDIX 4.20

SIGNIFICANCE TEST RESULTS FOR EFA

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
					Lower	Upper			
Pair 1	T1EFAq - T2EFAq	-5.625	5.772	1.443	-8.701	-2.549	-3.898	15	.001
Pair 2	M1EFAq - M2EFAq	-4.571	7.314	1.955	-8.794	-.348	-2.339	13	.036
Pair 3	T1EFAsub - T2EFAsub	-1.938	7.724	1.931	-6.053	2.178	-1.003	15	.332
Pair 4	M1EFAsub - M2EFAsub	.308	7.476	2.074	-4.210	4.826	.148	12	.885

Key: q – question level sub – sub-question

APPENDIX 4.2: IN-DEPTH INTERVIEW RESULTS

APPENDIX 4.21 TEACHERS' PERCEPTIONS

	Item	Perception	Percent respondents N = 13
1.	Usefulness of Peer instruction	Improving quality of test	46
		Improving test construction skills	31
		Improving instruction	15
		Improving test coverage (content)	8
		It provided practical approach to test construction	8
		Other teachers in school benefited too	8
		Originality in testing	8

	Item	Perception	Percent respondents N = 13
2.	Helpfulness of Peer instruction	Improved learner performance at MSCE	69
		Learners practiced more on better tests	39
		Reduced teachers reliance on past papers	31
		Encouraged students to work hard	31
		Assisted to follow learner progress	23
		Assisted to assess quality of items	8
		Assisted to handle examinations questions	8
3.	Reasons for using past examinations items	Lack of time for constructing test items	39
		Laziness	39
		To find out how learners can perform on MANEB items	31
		Lack of knowledge in test construction	23
		Scouting possible examination questions	8
4.	Challenges experienced	Overload in terms of periods per week	77
		Overload in terms of other responsibilities in school	57
		Overload in terms of class size (marking)	46
		Lack of text books and stationery	39
		Lack of chemicals and equipment for practical	39
		Negative attitude of school administration towards providing for science	8
		Overload in terms of nature of subject (very involving)	8
5.	Recommendations	Recruit more qualified Physical science teachers	77
		Motivate the Physical science teacher through frequent in-service training	69

	Item	Perception	Percent respondents N = 13
		Train more Physical science teachers	62
		Motivate the Physical science teacher through better salaries	54
		Encourage good teaching of Physical science	31
		Cover Physical science syllabuses on time in schools	23
		Motivate students to take and do well in science	23
		Supply adequate Physical science instructional resources to schools	23
		Motivate the Physical science teacher through promotion	8
		Motivate the Physical science teacher through paying risk allowances	8
		Motivate the Physical science teacher through paying allowances for preparing practical	8
		Motivate the Physical science teacher through employing a Laboratory assistance	8
		Motivate the Physical science teacher through provision of good accommodation	8

Note: Respondents were encouraged to say as much as possible.

AUTHOR RESUME

The Author was born on 11 May 1951 in FilimoniVillage, T. A. Simulemba in Kasungu. Between 1960 and 1973 he attended Chamakala, St Patricks Seminary, Kasina Seminary, Mzimba LEA and Chaminade Secondary Schools. He obtained a Diploma in Education and a Bachelor of Education Degree at Chancellor College in 1976 and 1984 respectively. In 1993 he obtained a Master of Education Degree (Science Education) at Makerere University in Kampala, Uganda.

The Author taught in secondary schools between 1976 and 1992. He was Headmaster of St John Bosco Secondary School from 1986 to 1989, St John's Secondary School in 1989 and Madisi Secondary School from 1990 to 1992. He was a Lecturer and Principal Lecturer at Domasi College of Education between 1993 and 1998. He served as a Senior Research and Test Development Officer and later as Principal Research and Test Development Officer for The Malawi National Examinations Board between 1998 and 2005. He served the Board as Director of Computer Services Department from 2006 to 2011 and retired.