Author	**NKPONE, Hycienth Leyiga**
	PG/Ph.D/96/22957
Title	**Application of Latent Trait Models in the Development and Standardization of a Physics Achievement Test for Senior Secondary Students**
Faculty	**Education**
Department	**Science Education**
Date	**October, 2001**
Signature	

APPLICATION OF LATENT TRAIT MODELS IN THE DEVELOPMENT AND STANDARDISATION OF A PHYSICS ACHIEVEMENT TEST FOR SENIOR SECONDARY STUDENTS

BY

NKPONE, HYCIENTH LEYIGA
PG/96/Ph.D/22957

BEING
A DOCTORAL DISSERTATION PRESENTED TO THE FACULTY OF EDUCATION, UNIVERSITY OF NIGERIA, NSUKKA

IN PARTIAL FULFILMENT OF THE REQUIREMENT FOR THE AWARD OF THE DEGREE OF DOCTOR OF PHILOSOPHY IN EDUCATION (MEASUREMENT & EVALUATION)

OCTOBER, 2001

CERTIFICATION

Nkpone, Hycienth Leyiga, a Doctoral student of the Sub-Department of Science Education of the University of Nigeria with registration number PG/96/Ph.D/22957 has satisfactorily completed the requirements for the degree of Doctor of Philosophy in Measurement and Evaluation.

The work embodied in this thesis is original and has not been submitted in part or whole for any other diploma or degree of this or any other University.

Dr. Uche Nzewi
Head of Department

Dr. B.G. Nworgu
Supervisor

APPROVAL PAGE

This thesis has been approved for the Faculty of Education,

University of Nigeria, Nsukka

By

Supervisor

Internal Examiner

Head of Department

External Examiner
22-10-2001.

Dean of Faculty

DEDICATION

This work is dedicated to my late mother, Mrs. Sarah Torbura Nkpone, and my son, Mr. Baribiea Leyiga.

ACKNOWLEDGEMENT

The researcher wishes to express his inestimable thanks of gratitude and appreciation to his Supervisor, Dr. B. G. Nworgu for exposing him to the areas of measurement theories, test models from where the study emerged. His guidance and expert advice equipped the researcher with the much-needed knowledge while this research was in progress. His accessibility and useful suggestions throughout the course of the study is profoundly acknowledged. The researcher is indebted to Drs. S. A. Ezeudu, F. A. Okwo, E. K. N. Nwagu, U.N.V. Agwagah, D.N. Ezeh, for their invaluable help and numerous insightful comments that guided the design and conduct of the study.

The researcher is equally indebted to Professor A.I. Joe for providing stimulating advice and encouragement and to Dr. R. P. I. Ukwuije for providing pertinent reference materials during the process of the study. The researcher wishes to thank Mr. Yellow Mazi of the Computer Science Department, Rivers State Polytechnic, Bori, who wrote the computer programs for data analysis, carried out some of the mathematical derivations and helped in other important aspects of the study. Special commendations also go to the Polytechnic management especially the Registrar, Mr. F. K. Puene for approving study leave for the study and to Messrs P. I. Igbara, E. N. Gerenwa, E.L.Digine, E. L. Nnaa and J. O. Alabi for their financial support and encouragement throughout the period of the study.

Special appreciation goes to Messrs B. G. Nsereka, S. A. Ngbo, Nekabari, N. A. W. Nkpone, T. Alugbara, N. B. Kpenu, Mrs B. C. Eze and Miss L. R. Odagme who provided the much desired succour and encouragement at crucial times. The principals of the sampled schools are most graciously acknowledged for allowing the researcher unhindered access to the subjects used in this research and the physics teachers who helped in collecting relevant data for the study. Finally, the researcher wishes to thank the Almighty God for the health and inspiration to carry out this study.

H. L. Nkpone
PG/96/Ph.D/22957

ABSTRACT

The development and validation of achievement tests have been founded on the classical test model. Unfortunately, the classical test model has been unable to solve adequately a number of testing problems during its history. Consequently, several alternative test models have been suggested. A frequently cited alternative to the classical test model is the latent trait model. The study therefore was aimed at developing and establishing the psychometric properties of a physics achievement test using the one-parameter, the two-parameter latent trait models and the classical test model. Altogether, 2215 male and female senior secondary school physics students participated in the study. The instrument for the study consisted of 60 multiple – choice items on a physics achievement test, developed by the researcher. Reliability and validity for each item and for the whole test were established according to the one-parameter and two-parameter latent trait models. Analysis was performed using PROX and regression techniques on the Microsoft Excel Visual BASIC computer programme. The chi-square goodness-of-fit test was used to investigate how well the one-parameter latent trait model fits the PAT items, while factor analysis was used to establish the unidimensionality of the final version of the PAT items. The results showed that the overall reliability co-efficient (KR-20) of the PAT was 0.89; the fit to the model was good, there was no significant relationship between easiness values obtained from two contrasting score groups. The physics achievement test measured one single trait for physics ability.

There was significant relationship, among the item parameters obtained from the one-parameter, the two parameter and the classical test models, the PAT item parameters and associated standard errors were reasonably estimated. The results obtained have implications, for latent trait investigation of item bias and for criterion referenced testing. One recommendation, based on the findings of the study, was that latent trait analysis should be followed up by some qualitative analysis. One factor which reduced the invariance of the item parameters and limit the generalizability of the findings of this study was that subjects with low scores were included in the validation process. Among others, it was suggested that a latent trait validation of criterion referenced test needed to be investigated.

TABLE OF CONTENTS

CHAPTER FIVE - DISCUSSION OF RESULTS, CONCLUSION, IMPLICATIONS AND RECOMMENDATIONS

APPENDICES

LIST OF TABLES

LIST OF FIGURES

CHAPTER ONE

INTRODUCTION TO THE STUDY

Background to the Problem

The development and standardization of classroom achievement tests have long been founded on the classical test model. This is primarily due to its simplicity both conceptually and computationally (Korashy, 1995: Smith, 1996). The classical test model views a score X , that a student obtains on an achievement test as the sum of two unobservable components, T and E, which form the relationship: X=T+E. X is the observed raw score; T is the true score, an increment: positive or negative, which is a function of conditions of test administration (Thorndike, 1990). The classical test model assumes that the best set of achievement test items is defined as a set of homogeneous items dominated by a single underlying dimension. To develop an achievement test, for example, a physics achievement test, using the classical test model, a large set of items are evaluated by selecting those items which yield a high point biserial correlation co-efficients.(Weiss & Davison, 1981: Thorndike, 1990: Ludlow & Bell, 1996).

Many workers such as Fraser (1980): Tobin & Capie (1982): Agwagah (1985): Nworgu (1985): Obioma (1985): Ogochukwu (1990): Ndukwe (1990): Nworgu & Harbor-Peters (1990) have based their studies on the classical test model which is no longer considered valid for ensuring objectivity in measurement (Korashy, 1995; Smith, 1996; Ndalichako & Rogers, 1997). The classical test model has been viewed as being too restrictive and inadequate in scope for the measurement required in evaluating the effectiveness of classroom achievement test (Izard & White, 1980). Thus the classical test model has some serious limitations that make its use in developing and establishing the psychometric properties of classroom achievement tests questionable.

One major limitation of the classical test model is the variant nature of the indices used to describe the item parameters (Douglas, 1990: Guiton & Ironson, 1983). The proportion of examinees in a sample who get an item correct (Item difficulty) changes from a sample whose mean ability is high to one whose mean ability is low. Consequently, the same individual tested in two different samples may obtain two different errors of measurement and estimate of true score (Weiss & Davison, 1981). Item parameters or statistics which remain invariant or stable from one sample of examinees to another have been viewed as a desideratum for any psychometric measurement (Douglas, 1990). Gulliksen (1950: 392) noted this basic limitation of the classical test model and states that:

> A significant contribution to item analysis theory would be the discovery of item parameters that remained relatively stable as the item analysis group changed or the discovery of a law relating the changes in item parameter to changes in the group.

Gulliksen identified invariant item parameters as an outstanding measurement problem and suggested that whosoever elucidated it would have made a significant contribution to item analysis theory. Lord and Novick (1968) have also expressed a similar opinion. In addition, the sample-based nature of all the estimation procedures used to establish reliability co-efficients is another weakness of the classical test model. The reliability estimates are specifically a function of a particular set of items composed of a test and sample of examinees on which that data were collected. This means that in the classical test model, the contribution of each item to the test reliability and validity depends upon what other items are in the test. Furthermore, the classical test model typically provides only one overall index of standard error of measurement, the most rudimentary indication of reliability in the classical test model, for all items on the entire test (Guiton and Ironson, 1983). As Guiton and Ironson (1983) point out, it is unreasonable to assume that scores throughout an entire test will have the same degree of precision of measurement or one overall index of dependability. To this end, the classical test model is no longer considered adequate in developing and establishing the psychometric properties of achievement tests.

Again, scores from test based on the classical test model are interpreted in terms of percentile ranks or as the standard deviations from the mean of scores in the standardizing sample (Thorndike, 1990). A different test or a different sample results in a different unit. Consequently, the classical test model does not ensure objectivity in measurement, as one technique of measuring a particular attribute is not interchangeable with an alternative measure of the same attribute unless both are interpreted in terms of common group-norm. Objective measurement ideally yields scores or attributes that are interpreted with direct reference to the attribute being measured and one technique of measuring a particular attribute is interchangeable with an alternative measure of the same attribute without reference to the standardization group (Guiton & Ironson, 1983: Thorndike, 1990).

Another limitation of the classical test model is that the total raw score obtained by an examinee on an achievement test does not take into account the examinee's answer patterns on the test. That is, the model ignores the patterns of item responses. Such information is desirable in classroom instruction, for instance when the teacher wants to ascertain strategies the examinee used to answer a given item. In addition, classical test model provides no basis for determining what a particular examinee might do when confronted with difficult test items. According to Hambleton and Cook (1977), such information is necessary if the test developer desires to predict test scores characteristics in a subgroup of examinees. Another deficiency of the classical test model is that because observed scores contain error, one is not quite sure of the psychometric properties of the achievement test that is being estimated and on what basis the estimation was done (Wilcox, 1979).

Under the classical test model, scores on achievement tests, for instance, a physics achievement test, is often expressed as the total raw score the examinee obtained on the content areas concerned. This form of assessment focuses the teacher's attention on the total scores and on ranking the students rather than on ascertaining why a student did not succeed on a particular test item. Thus teachers tend to use achievement on a test based on this model to label and categorise students (Keats, 1980). In addition, the classical test model's development technique does not ensure that two or more forms of an achievement test can be made equivalent in level and range of difficulty (Izard & White, 1980). Again, using the classical measurement model does not ensure or allow meaningful comparison of examinees who have been tested by a non-parallel achievement test. For instance if a physics achievement test is substantially easier or it taps a wider range of ability, the scores obtained with it are not comparable to scores on another one. As Guiton & Ironson, (1983) noted, in testing situations it is often desirable to compare achievements of examinees who have been tested with different set of items. This restriction is because the classical test model is based on the conception of exactly parallel tests, an assumption that is rarely satisfied (Guiton & Ironson, 1983).

Because the classical test model has been unable to solve adequately a number of testing problems during its history, several alternative test models, such as criterion referenced test; nominal response model and graded response model, have been suggested (Samejima, 1969; Bock, 1972; Susan, 1983). A frequently cited alternative to the classical test model is the latent trait model (Korasy, 1995; Smith, 1996). The latent trait measurement model is a mathematical function relating the probability of success on a cognitive task to the underlying ability or other characteristics measured by the task (Bentler, 1980, Gallini, 1983).

A latent trait model specifies a relation between the empirical test performance and unobservable trait underlying the test performance. Basically, latent trait model is concerned with how the probability of success on a test item varies as a function of ability (θ) and other parameters represented by item difficulty, discrimination power and the degree of guessing (Baker, 1977). The theory of latent trait states that in testing situations, examinee's performance on a test can be predicted or explained by defining the examinee's characteristics referred to as traits and using the scores to predict or explain test performance (Lord & NovicK, 1968).

Latent trait does not mean that the trait is static. It refers to characteristics of examinees that are not directly measurable or observable but are assumed to underlie test performance (Guiton & Ironson, 1983). The latent trait is referred generically to as ability (θ) which the test item is attempting to measure, latent trait are not directly measurable or observable hence the proportion of examinees passing on the test are used as estimator of trait values (Andrich, 1980; Wainer, Morgan & Gustsfassion, 1980). These proportions are the manifest variables. Therefore, a latent trait model fits a set of data if it correctly predicts the manifest variable (Cressie, 1983).

The basic mathematical linear form of the latent trait model is expressed as follows: $Pg.(\theta) = bg + ag(\theta)$; $g = 1 \ldots n$, designates the probability that an examinee with a given latent trait , ability (() will answer item g (g= 1,2,3,4........n) and n is the number of items. The ag parameter is the item discriminating index; bg is the index of item difficulty (Gallini, 1983). The basic mechanism for determining a latent trait test model lies in the item characteristic curve, which is a graphical display of the relationship between the probability of success on an item and the examinee's position on the underlying trait measured by the test (Wainer, Morgan & Gustsfsson, 1980). The item characteristic curve is defined as a mathematical function that relates the probability of success on an item and the examinee's position on the underlying trait measured by the test (Wainer, Morgan & Gustsfsson, 1980).

The item characteristic curve is defined as a mathematical function that relates the probability of success on an item to the ability (θ) measured by the test (Birnbaun, 1968). This is expressed as:

$Pg = F(\theta)$. Since scores on the underlying trait are generally not observable or measurable, observed scores are used as estimator of trait values (Sax, 1989; Reiser, 1983).

Latent trait models are usually differentiated by the number of parameters estimated from the item. For instance, the one-parameter model, also known as the Rasch model (Rasch, 1966) is a special case of the latent trait models. In this model, the parameter for discrimination power is assumed to be the same for all the items. This means that in the one-parameter model, test items are described only in terms of their difficulties. The two-parameter model describes item by both their difficulties and discrimination indices. The latent trait model contains person parameter, ability estimate (θ) which can be used to measure individual differences. The prototype method of assessing learning outcomes in classroom instruction based on the classical test model ignored this very important factor in learning (Gagne, 1988). Thus the latent trait model is the only measurement model that takes individual differences into consideration when evaluating cognitive learning outcomes (Susan - Embretson, 1983).

The latent trait models generally satisfies Gulliksen's requirement because the estimate of the achievement test item parameters and person's ability does not depend either on which subgroups the person belongs to nor on the selection of the specific set of items, provided the data fit the model (Wainer, Morgan & Gustafsson, 1980). The Latent trait model is the only measure techniques that possess invariant item parameters. This is because in the latent trait model, the function of a test is concerns with estimation of an individual location on a dimension represented by a trait rather than being expressed in normative terms. The item parameters are therefore invariant with respect to the ability of subjects in the validating sample (Lord & Novick, 1968, Creassie, 1983).

Another explanation for the invariant item parameters of the latent trait measurement model as against the classical test model is that in the classical test model, item statistics are computed from a sample with a distribution of ability like that in the population as a whole (Guiton & Ironson, 1983). In the latent trait item analysis, the ability distribution in the research sample may have different distribution from that of the population. As Guiton & Ironson (1983) note, this relaxation contributes to parameters invariance in the latent trait model.

The major potential of latent trait models rests in the characteristics of parameter invariance. Item parameters are invariant across groups of examinees and ability parameters are invariant across groups of items. The invariance of the item parameters is particularly important in horizontal equating settings, while the invariance of ability parameters is essential in latent trait applications in vertical equating and computerised adaptive testing (Allen, Ansley & Forsyth, 1987). In horizontal equating, the equating process adjusts for unintended differences in difficulty levels or differences in ability score distribution. In vertical equating, the test forms are developed to differ in difficulty levels and the groups of examinees involved in the equating process also differ in ability level (Angoff, 1971).

It is the property of invariant ability parameters that has enabled users of standardised achievement tests to consider the development of tailored tests which use only those items in a test that fit the ability of the learner. The benefits obtained as a result of the invariance of the parameters of the latent trait model are achieved only when certain conditions are satisfied (Baker, 1977; Wainer, Mogan & Gustafsson, 1980; Allen, Ansley & Forsyth, 1987):

1. The test items must be uni-dimensional; the test items must measure a single trait.
2. The test should be equally reliable for all subgroups of examinees
3. The trait must be sufficiently defined so that the test items can be developed to measure the trait.
4. There must be local stochastic independence among the items.

Another major advantage of latent trait models in the development and standardisation of classroom achievement tests is that they provide not only an index of standard error for scores on the whole test but also an index of standard error for each person ability and each item. In latent trait analysis, standard errors are expressed in logits. The logits are interval level units of measurement corresponding to the total raw scores that have undergone exponential transformations. To this end, inferences and validation on achievement tests made using these models are, therefore, derived from a solid statistical basis (Korashy. 1995). The estimate of standard error for each test item and person (raw score) makes it possible to select achievement test items that will make maximum contributions to the test design and the measurement efficiency at a particular ability level. Thus, at each ability level, a more precise confidence statement can be made than the conventional one overall standard error of measurement for the total test computed in the classical test model (Guiton & Ironson, 1983).

Applying latent trait models in the development and standardization of classroom achievement tests, such as a physics achievement tests, make it possible to analyse the pattern of each examinee's answer to the test through analysis of person fit to the model. If the fit statistics for an examinee's achievement on the test is acceptable, then the examinee's achievement is taken as a valid basis for inferring a measure of the examinee ability. Knowledge of such response patterns also have implication in counselling (Baker, 1977). The latent trait measurement models handle test scores equating problem in the classical test model using ability estimates and in doing so, they are able to control for problem of non-equivalent groups (Morgan, 1980). The ability estimates in the latent trait model can also be used to measure individual differences in classroom instruction. The latent trait model is thus the only measurement model, that takes individual differences into consideration in testing situations (Susan, 1983). Another major advantage of the latent trait models is its objectivity. This objectivity stems from the fact that the comparison of ability does not depend on which particular group of examinees used to validate the item parameters or on a specific set of items (Tinsley & Dawis, 1975; Douglas, 1980). Furthermore, latent trait measurement models have recently attracted considerable attention from measurement specialists as a result of a successful application of the theory to such problem areas as conceptualising test items, equating different forms of the same test, presenting items tailored to individual examinee and the precise determination of fit by the use of the chi-square statistic (Hambleton & Cook, 1977; Wright, 1977; and Wainer, Morgan & Gustafsson, 1980).

One important difference between developing classroom achievement tests using the classical test model and models based on latent trait occurs during selection of test item and item validation. Item validation techniques involve the use of standard error of measurement, the statistics of maximum likelihood and the analysis of fit for revising the test items. Unlike the classical item analysis, where desirable characteristics of a test are high reliability and item validity, in the latent trait item analysis, the essential criterion is the compatibility of the items with the model, that is, the fit of the items to the model (Smith, 1996). The fit implies validity. Thus, while the classical item analysis is conceptually easy to understand and is manually calculable, the flaws embedded within it make its use in developing and establishing the psychometric properties of classroom achievement tests questionable.

Although the assumptions and mathematics of latent trait models are more complex, costly and timeconsuming than the classical test model, several authors such as Smith (1996) and Korashy (1995), have argued that their practical advantages are sufficient to warrant their usage. Again, a closer examination of the literature reveals that there is no empirical study on latent trait measurement models using a physics achievement test. These arguments in favour of latent trait models are quite convincing to justify their applications in the development and standardisation of a physics achievement test for Senior Secondary Physics students.

Evaluation of the effectiveness of teaching and learning has raised a number of questions concerning the psychometric properties of classroom achievement tests. As Izard and White (1980) note, the essential problem in evaluating the effectiveness of teaching and a broadly conceived educational goal is one of validation. This essentially boils down to the question of whether a given achievement test, such as a physics achievement test, provides a satisfactory psychometric properties of the concept it was designed to measure. Therefore, in order to evaluate cognitive outcomes in classroom instruction, consideration must first be given to the problems involved in the development and standardization of the achievement tests.

Among the various factors that might be responsible for the poor performance in achievement testing, it is strongly suspected that poor methods of evaluating the tests would contribute immensely. This is because evaluation itself plays a vital role in making teaching and learning effective (Nworgu, 1992: Mehrens & Lehman, 1978). For instance, carefully collected evaluation data on an achievement test, such as a physics achievement test, help the physics teacher to understand the learner, plan learning experience for him and determine the extent to which the instructional objectives are being achieved (Grondlnd, 1980). Therefore, if the evaluation instruments, the achievement tests, are faulty, inappropriate decisions on various issues affecting the students might result. Instructional decisions are more likely to be sound when they are based on information that is accurate, relevant and comprehensive. One of the best ways to ensure accurate decision is to utilise a valid and reliable achievement tests (Choppin, 1980). The adoption of continuous assessment of learning outcomes in secondary and tertiary institutions in Nigeria (Federal Ministry of Education, Science and Technology, 1985) amplifies further the need to develop and validate relevant tools for measuring cognitive learning outcomes in classroom instruction. In developing achievement test, it is a common practice to use a great number of items. The question of how the items may be sufficiently truthful, genuine and accurate is of prime consideration (Choppin, 1980).

When there are several levels of a given achievement test, one of the major problems is making sure that scores on one test is comparable so that when one version of the test composed of one set of items is administered later, differences in scores can be attributed to changes taking place in the individual tested rather than in the test itself. The extent to which test scores can be equated, that is, placing the test scores on the same scale so that test items of appropriate difficulty are adapted for each examinee, depends on the psychometric properties of the test (Angoff, 1971: Morgan, 1980). To this end, in order to develop parallel achievement tests and equate the test scores, the psychometric properties of the test must be given due consideration.

Furthermore, the issue of bias in achievement test items (Smith, 1996) has necessitated further the need for establishing the psychometric properties of an achievement test /or of obtaining quantitative information pertaining to the test items, when investigating item bias. Smith (1996) defines bias in achievement test items as differential validity of a given interpretation of a test score for any definable subgroup of examinees. Consequently, in testing and evaluating significant item X group interaction in achievement test items, prime consideration must be given to the problem involved in the development and standardization of the achievement test, such as a physics achievement test.

Achievement tests measure the present proficiency, mastery, and understanding of general and specific areas of knowledge (Kerlinger, 1973). They are measures of the effectiveness of instruction and learning. Achievement tests are important in education and educational research. Specifically in research involving instructional methods, achievement is often the dependent (criterion) variable.

Achievement tests attempt to measure what an individual has learned and his present level of performance or accomplishment. The assessment of terminal or criterion behaviour involves the determination of the psychometric characteristics of the achievement tests as well as the evaluation of the student's achievement (Bloom, 1976). Ebel (1972:257) gives the following opinion about academic achievement:

> Academic achievement is the intellectual competence of human beings; to be important, educational achievement must lead to a difference in behaviour. Broadly conceived, academic achievement is the general term for the successful attainment of some goals requiring a certain amount of effort and the degree of success obtained in an intellectual task.

Achievement tests are designed to measure the outcome or level of accomplishment in a specified instruction which a student had undertaken and are used to find out how much of the material a student has learned (Nkemakolam, 1992). Achievement has always been associated with such concepts as objectivity, accountability, evaluation and grades (Izard & White, 1980). Achievement tests are used in evaluating the influence of courses of study, on the learner teaching methods, diagnosing strengths and weakness in classroom instruction and as a basis for decision making (Best, 1981).

Statement of Problem

The development and standardization of classroom achievement tests, for example, a physics achievement test, has, for long, been founded on the classical test model. Unfortunately, the classical test model has some serious disadvantages which made its use in developing and establishing the psychometric properties of classroom achievement tests untenable. One major limitation of the classical test model is the variant nature of the indices used to describe the item parameters. Furthermore, the classical test model typically provides only one overall index of reliability for all items composed of a test. In addition, the sample-based nature of all the estimation procedures used to establish reliability co-efficients is another weakness of the classical test model. The classical test model does not ensure objectivity in measurement, as one technique of measuring a particular attribute is not interchangeable with an alternative measure of the same attribute.

Besides, the classical test model has failed to provide satisfactory solutions to many testing problems such as test score equating, test design and testing items for bias. For these and other reasons, psychometricians have been investigating and developing more appropriate test theories. Consequently, considerable attention is being directed towards the field of the latent trait model. Advantages of the latent trait model in providing invariant item parameters, conceptualizing classroom achievement tests, equating different forms of the same test, presenting test items tailored to individual examinees, testing test items for bias and the precise determination of the fit make it worthwhile to explore the application of latent trait models in the development and standardization of classroom achievement tests, such as a physics achievement test. In addition, in latent trait analysis, each item parameter is accompanied by its standard error of measurement, the most rudimentary indicator of reliability in the classical test model. To this end, inferences and validation of achievement tests based on the latent trait models are derived from a solid statistical basis. Advantages of latent trait models make it justifiable in considering their application as a promising addition to the repertoire of measurement methods in the development and standardization of classroom achievement tests.

When evaluating the effectiveness of teaching and learning, equating test scores, developing parallel forms of a given achievement test, for example, a physics achievement test, testing for individual differences in cognitive learning outcomes and testing items for bias, it is highly desirable to take into consideration some quantitative information regarding the psychometric properties of each item to be included in the final form of the test. Therefore, the problem of the study is: What are the psychometric properties of a physics achievement test developed and standardized for senior secondary school physics students using the one-and two-parameter latent trait models; and the classical test model; and the influence of each of the models on the item validity indices?

Purpose of the Study

The primary purpose of the study was to develop and establish the psychometric properties of a physics achievement test using the one-parameter and the two parameter latent trait model . Specifically, the study sought to:

1. Investigate the unidimensionality, single trait for ability of the physics achievement test.
2. estimated the item parameter, item difficulty indices using the one-parameter model.
3. estimate the item parameters, difficulty and discrimination indices using the two-parameter latent trait model.
4. estimate the subject measures, ability estimates and its associated standard errors.
5. estimate the standard error for each of the physics achievement test item.
6. investigate the fit of the physics achievement test items to the one-parameter latent trait model.

7. estimate the classical test model item parameter, point biserial correlation co-efficients, in order to provide a baseline against which latent trait models item parameters were compared.
8. investigated the invariance of the latent trait item difficulties with respect to the ability of the standardizing sample.
9. investigate the invariance of the ability estimates with respect to the ability of the standardizing sample.
10. investigate the relationship among the one-parameter, the two-parameter and the classical item point biserial correlation co-efficient.
11. investigate the relationship between the classical raw scores and its logit exponential transformation.

Scope of the Study

The study was delimited to two latent trait models. They were the one-parameter and two-parameter latent trait measurement models. These models were selected because of their frequent use in the research literature and because of their mathematical tractability (Reckase, 1979). This means that these models have the advantages of greater computational simplicity and interpretability. The analysis of models fit was delimited to the one-parameter model, because it is only in the one-parameter latent trait model that the raw score is a sufficient statistic or estimator of the latent ability.

The content of the physics achievement test was based on: I. Matter, position, motion and time; ii. Energy (mechanical and heat); iii. Waves. Iv. Fields. V. Atomic/nuclear physics and electronics. The physics items were delimited to Bloom's (1972) taxonomy of the cognitive domain educational objectives.

Research Questions

The study was carried out through answering the following research questions.

1. How uni-dimensional, single trait for ability are the physics achievement test items?
2. What are the estimates of the item parameter (difficulty indices) using the one-parameter model.?
3. What are the estimates of item parameters (difficulty and discrimination indices), using the two-parameter latent trait model?
4. What are the estimates of the subject measures, ability estimates and their associated standard errors indicated by the raw scores on the physics achievement test?
5. What are the estimates of the standard error (reliability) for each of the physics achievement test items?
6. To what extent do the physics achievement test items fit the one-parameter latent trait model?
7. What are the estimates of the classical test model item parameters, difficulty indices and point biserial correlation co-efficients?.

Hypotheses

The following null hypotheses were formulated and tested at 0.05 probability level.

1. There will be no significant relationship between the latent trait model item difficulties obtained from two contrasting ability groups of the same standardisation sample ($P< 0.05$).
2. There will be no significant relationship between the ability estimates obtained from two contrasting score groups of the same standardisation sample ($P <.0.05$).

3. There will be no significant relationship among the item parameters obtained from the one-parameter model, the two-parameter model and the classical item point biserial correlation co-efficient ($P< 0.05$).
4. There will be no significant relationship between the classical raw scores and its logit exponential transformation ($P<0.05$).

Assumptions of the Study

The study assumes that ability in term of the physics achievement test is normally distributed over the whole population from which the sample was drawn. This supposition is an indication of the order validity of the PAT and also a measure of a single trait for physics ability. The study further assumes a local independence of items and examinees. This is one of the basic requirements of the latent trait models. This assumption presupposes that the estimation of the item parameters is independent of the specific content areas or subgroups of examinees on which the validation was made. The researcher further assumes as follows:

1 Speed does not influence the probability of answering an item correctly. This assumption is in view of the fact that item analysis cannot be reasonably conducted in a test in which no examinee has time to complete all items.
2. Guessing is irrelevant as the study was based on the one-parameter and the two-parameter latent trait models, in which the guessing parameter is assumed to be zero.

Significance of the Study

The study will be of significance to test developers, practising teachers, counsellors and test users in the following areas:

The study will be of significance to test developers and practising teachers in the development of parallel achievement test and test scores equating due to the invariant properties of the latent trait model item parameters. Latent trait models allow for equating test scores using ability estimate and in doing so, they are able to control the problem of non-equivalent groups in the classical test model.

Test developers will find the study very useful in the determination of the presence or absence of item bias. A test is unbiased if all individuals having the same latent ability have equal probability of getting the item correct. Presence or absence of bias is readily assessed applying the latent trait model using either the item characteristic curves approach or analysis of residuals between subgroups. Due to its sample-invariant property, the latent trait model appears to be the only sound theoretical approach to the problem of test item bias detection.

It will provide practising teachers with the knowledge of tailored testing in the selection of test items, so that items of appropriate difficulty are chosen for each examinee. Validating classroom achievement tests using latent trait models results in logarithmic ability estimates being assigned to every raw score. This estimate indicates the amount of ability required to achieve a given raw score so that the items can be tailored to each examinee's ability.

The use of item fit will enable the test developer to choose different content areas of a given achievement test and to use them to measure traits on a common content. The fit analysis will also have implication for tailored, adaptive testing.

The study will not only provide test developers and practising teachers with the knowledge and capacity of estimating examinee abilities, even when some items have not been answered (that is, total scores from partial scores), it will also provide solutions to many practical testing problems such as the evaluation of essay writing ability, the use of graded scoring model and the analysis of fit by the use of chi-square statistic.

The study will make item tryouts more economical and informative for test developers and practising teachers. It will also permit the estimation of item parameters from any group of examinees thus making sample item tryouts unnecessary.

The study will permit the identification of each examinee's strength and weakness in testing situations and to use this diagnosis to find what the examinee needs next. This will be possible through the analysis of examinee's fit to the model and the evaluation of each examinee's answer patterns on an achievement test.

It will provide practising teachers with various methods of approximating classical item parameters into their latent trait equivalence in test development and item validation.

Test developers will find the study very useful in the development of criterion – referenced test which is geared towards assigning examinees to mastery state, due to the invariant property of the latent trait item parameters.

The study will have implication for counselling since information can be obtained on each examinee's strengths and weaknesses in a testing situation.

The study further has the following theoretical significance:

It will provide a firmer foundation for the evaluation of aberrant items or items for fidelity. Such items function differently in differing ability groups.

The analysis can be used to develop attitude scale items involving an ordered response categories. This implies that the study has the advantage of providing a precise measurement at each level of the attitude trait continuum.

In tailored measurements where the emphasis of testing is on obtaining different contents of items that can be used interchangeably without the classical requirement of parallel forms test.

Finally, recommendations at the end of the study will provide bases for further research in this area of achievement testing.

CHAPTER TWO

REVIEW OF RELATED LITERATURE

This chapter is devoted to the review of related literature and is organised under theoretical framework, empirical studies and summary of literature review.

1 Theoretical Framework

- (I) Historical perspective to achievement testing and measurement
- (ii) Historical and conceptual background of the classical test model
- (iii) Historical and conceptual background of latent trait models
- (iv) Item analysis
  - (a) Classical item analysis
  - (b) Latent trait item analysis
- (v) Connection of latent trait item parameters with the classical item statistics
- (vi) Reliability and standard error of measurement
- (vii) Validity and analysis of fit

2 Empirical Studies associated with Instrumentation Research

- (I) Studies on instrumentation research based on the classical test model.
- (ii) Studies on instrumentation research using latent trait models

3 Summary of literature Review

1. Theoretical Framework

i. Historical perspective to achievement testing and measurement.

Oral methods of assessing achievement have been in existence since the history of mankind. In the pre Christian time, the Ephesians and the Greeks used oral test to assess abilities of members of their societies (Ohuche & Akeju, 1977). In West Africa, indigenous education was mainly oral and informal. This system of education was also associated with the evaluation of achievement which was mainly oral and based on careful observation (Fafunwa, 1974).

DUBOIS (1970) attributes the increase in the use of testing and measurement to three major areas of development namely; Civil Service exams, School exams, and the study of individual differences. Civil service testing began in China about 300 years ago when the emperor decided to assess the competency of his officials. Though that system of testing was formal and explicit, the learning was informal in nature. Later, government position were filled by persons who scored high in examinations that covered such topics as music, horse-craftsmanship, civil law, writing, Confucian principles and knowledge of public and private economics (Ohuche & Akeju, 1977).

Paradoxically, as the Chinese were phasing out their exams, civil service exams were being developed in Britain and the United States as a fair way of selecting among job applicants for government jobs. Early evidence of the effectiveness of the exams was anecdotal in nature, although the exams were popular because they removed decisions from biases of political judgements. According to Dubois (1970), oral testing continued to dominate the mode of evaluating achievement as closer to our time as when the great Universities in

Europe introduced their first oral examination for Doctor of philosophy Degree at the University of Bulgaria in 1200 A.D.

Students in European schools were given oral exams until well after the 12th century when paper began replacing parchment and papyrus. In the 16th century, the Jesuits started using tests for the evaluation and placement of their students. Present-day students owe their exams to the later development of printing processes and allied industries. By the 18th century, the default of the oral testing had become widely recognised as it was not only unreliable, but was also affected by confounding factors such as personality, interactions among examiners and examinees (Ohuche & Akeju, 1977).

In contrast to the informal testing identified with the imperial China, modern testing allows for general competence required for a greater population and for the effectiveness of the system (Scates, 1947). Besides these competencies, modern testing is identified with formal learning. It is also geared toward the effectiveness of the system of education.

In the United States of America, formal examination is recognised as a means of raising standard in the State system of education. According to Dubois (1970), the study of individual differences began in Great Britain when Sir Francis Galton (1822-1911) set up his famous Anthropometrics laboratory containing instruments to measure various sensory and motor skills. In 1888, visitors at the international Health exhibition centre in London paid three pence each to be measured by Galton's test. Another Englishman, Karl Pearson (1857-1936), the founder of statistics developed a number of statistical techniques that forms the core of basic measurements theory. These include; the well known Pearson-product moment correlation co-efficient and the chi-square goodness-of-fit test.

In France, Alfred Binent (1857-1911) developed the first individual tests of intelligence (1905) as part of his work on the study of individual differences. Back in Great Britain, Charles Spearman (1863-1945), developed our modern concepts of test reliability and factor analysis. Later William Stern, in Germany, got the notion that in judging people's potential, it was not their mental age alone but their mental age in relation to their chronological age (CA) that was important. He suggested dividing mental age by chronological age to obtain a quotient that, when multiplied by 100, she called the Intelligence Quotient. According to Dubois (1970), measurement theory as a discipline began to blossom in the 1930's. In 1935, the journal, Psychometrika was founded. Educational and psychological measurement followed in 1941. The British Journal of statistical psychology began in 1947 and Multivariate Behavioural Research began in 1966. The first textbook in measurement theory, E. L Thorndike, "An Introduction to the Theory of Mental and social Measurement appeared in 1904.

Back in West African sub-region, Western education was introduced at the colonisation of the territory. Specifically, the first Secondary School was established in Nigeria in 1859 by the Church Missionary Society (Adesina, 1984). Associated with this new system of education were Western methods of educational evaluation which were at the beginning of the 19th century under going transformations in the United Kingdom and United States of America. With the progress of Secondary education in the period that followed, it became necessary to establish the West African Examinations Council (WAEC). WAEC was established in 1952 with the sole purpose of conducting Examination in the public interest (Ezeobar, 1983). The council was also to award certificates, provided these certificates were of comparable standard to certificate of examining bodies in the United Kingdom (Ohuche & Akeju, 1977).

In 1966 the Test Development and Research Unit (TEDRU), an organ of WAEC was established. The Test Development and Research unit develops standardised tests for the council. Mention must also be made of a significant event that took place in the history of achievement testing in Nigeria and Africa in establishing the International Centre for Educational Evaluation (ICEE) at the University of Ibadan in 1972. The centre resulted from a joint Co-operation among the University of Ibadan, the Science Education Programme for Africa, and the Carnegie Corporation at New York (Ohuche & Akeju, 1977). The functions of the centre according to Ohuche & Akeju are to: train educators, undertake research involving evaluation instruments, and provide talented personnel for the evaluation of specific personnel for the evaluation of specific projects.

The Joint Admission and Matriculation Board (JAMB) was established in 1978 to cater for admission of candidates into Nigerian Universities. Recently, in 1982, the Nigerian government adopted a new method of assessment, the continuous assessment as against the former terminal examination at the end of each level of education (Federal Ministry of Education, Science and Technology, 1985). Achievement measurement is distinguished from aptitude measurement in that the instrument used to assess achievement is specifically concerned with the characteristics and properties of present performance with emphasis on the meaningfulness of the content, while aptitude measurements are aimed at predicting students later performance in a specific type of behaviour and are used essentially in educational research to equate initial groups that are to receive different experimental treatment (Best, 1981; Borg & Gall, 1983). According to Bloom (1976) three factors determine a student academic achievement.

1. Cognitive entry behaviour, that is the extent to which the student has already learned the basic pre-requisite for the learning to be accomplished. Since

the amount of time needed to learn is obviously a function of how much has been learned already, these pre-requisites constitute a necessary link between the learner and the accomplishment of the learning task.

(2) Affective entry behaviour, that is, the extent to which the student is or can be engaged in the learning process. Affective entry behaviour is generally expressed in attitudes, interests, self-concepts and emotional sets.

(3) Quality of instruction, that is the extent to which the instructional cues are appropriate for each student, the amount of participation, practice of the learning task by each student and the type of reinforcement given to each learner. Related to these variables are: the variety of instructional models and materials, the teacher verbal ability and the type of feedback available to both the teacher and student. To him, the quality of instruction is a major determinant of who will learn well. Bloom (1976) assumes that the three variables, taken together, account for as much as 90 percent of achievement variance. Although most of the foundation for the classical measurement theory had been completed, Psychometricians are currently developing alternative test models and techniques, one of which is the latent trait model.

ii. **<u>Historical and conceptual background of the classical test model.</u>**

The classical test model has its origin in physical measurement and the study of errors of measurement (Keats 1980). The model assumes that a personscore (Xi) on an achievement test may be regarded as consisting of two additive components: the true score (Ti) and error score (Ei) so that: $Xi = Ti + Ei$

The true score (Ti) is thought of as a parameter of the person at the time of testing, while the error score (Ei) is thought of as due to minor fluctuations caused by irrelevant factors (Guilford, 1954). The true score is also considered as the average score of an individual examinee in repeated measurement, while the error score is an increment, positive or negative which is a function of conditions at a particular occasion of test administration to a particular examinee (Guiford 1954). The basic assumptions of this model are (Keats, 1980: Thorndike, 1990):

(1). The obtained score is the sum of true score plus error score, that is:

$X_{obs} = X_{true} + X_{error}$

(2) Over the population, error score is independent of true score.

(3) In pairs of measures, the error in one measure is independent of the error in the other.

(4) For a given population of examinees repeatedly tested on a number of occasions, error variance is assumed constant over repeated testing.

Given these assumptions, Guilford (1954) contends that as the number of persons or the number of measurement increases, the mean error approaches zero as a limit. This follows from the definition of errors as random deviation from the true score that are generally positive or negative. Also, in the limit the mean of observed score is equal to the mean of true score, that is:

$$\overline{X}\text{ observed} = \overline{X}\text{ true}$$

The true score and error score were conceived of as completely independent. The true score was viewed as unchanging from one form of test to a parallel alternative form, and from one occasion of test administration to another. The error was considered to be unique to the specific measurement, and to be entirely independent of the error expected to appear on another measurement designed to assess the same characteristic of a person (Keats, 1980; Thorndike, 1990). The true score is unobservable. Its existence and properties could only be inferred from consistency of performance from one test administration to another or from one test score to another.

The classical test model with its concept of reliability assumes that the best set of items is a set of homogeneous items dominated by a single underlying factor (Weiss & Davisson, 1981). This model of true and error scores was conceived of as a productive model, which led to the formulation of a number of useful relationships. It produced a statement of the relationship between test precision and test length; it permitted the development of estimates of precisions of difference scores and chance scores. In addition, the classical test model led to estimates of the degree to which indices of relationship between measures of difference attributes are attenuated by error of measurement (Keats, 1980).

This classical test model yields an estimate of reliability from the product moment correlation, between observed measures from any two occasions of measurement (Allen & Wendy, 1979). From the empirical point of view, the applicability of this model rests mainly on the supposition that the true score for any given individual remains constant over occasions of measurement (Kim, 1979).

iii. Historical and conceptual background of latent trait models.

The term "latent trait" was first introduced by Lazarsfeld (1950) and the work of Lord (1953) gave "birth" to latent trait test models. The latent trait represents ability measured by the test. Latent traits are psychological dimensions necessary for the psychological description of individuals (Lord & Novick, 1968). This means that latent trait is a theoretical attribute or characteristics that independently and completely defines test performance. According to Cressie (1983), the development of latent trait models begins with assumptions that the probability that a person will answer an item correctly is the product of an ability parameter, pertaining to the person and the difficulty parameter, pertaining only to the item.

However, the ability of the examinee is not directly observable; it is referred to as latent traits or latent variables. Latent trait does not mean that the trait is static. It refers simply to characteristics of examinees that are not directly measurable or observable but are assumed to determine test performance, and they are referred generically to as ability (θ) which the test item is attempting to measure (Guiton & Ironson, 1983). Recent research in latent trait model is confirmed largely to the normal ogive model which was elaborated by Lawley (1943) and Lord (1953) and the logistic models by Rasch (1960: 1966) and Birnbaun (1968). Since there is a high degree of similarity between the two (Baker, 1977) the logistic ogive is used in this study. In contrast to the classical test model, the latent trait model makes strong assumptions about individual behaviour when answering a test item (Allen and Wendy, 1979). Lazarsfeld (1950), developed a theory in the context of attitude tests and other short questionnaires. Lord (1953) worked on the theory of aptitude and achievement test. Birnbaun (1968) published a series of reports using item characteristic curve models. The use of item characteristic curve functions was initiated by Ferguson (1942) and elaborated by Lawley (1943); Tucker (1948); Lord (1953) and Lord and Novick (1968).

The graded response model was introduced by Samejima (1969) to handle testing situations where item responses are made into two or more ordered categories such as incorrect, partially correct and correct. The graded response model assumed that any response to an item g can be classified into m+l categories, scored Xg = 0, 1 ..... mg respectively. Samejima (1969) introduced the operating characteristic of a graded response category which she defined as:

$$Px_g\,(\theta) = P_{xg}\,\theta - P_{(xg+1)}\,(\theta)$$

Where P_{xg} ((θ)) is the regression of the binary item score on latent ability, when all the response categories less than X_g are scored 0, and those equal to or greater than X_g are scored 1. Pxg(θ) represents the probability with which an examinee with ability level (θ) receives a score of xg.

In view of the fact that the one –, two and three parameter models can only be applied to test items that are scored dichotomously, the nominal response model was introduced by Bock (1972) and elaborated by Samejima (1972). The nominal response model is applicable when items are multichotomously scored (Bock, 1972). The purpose of the model is to maximise the precision of the obtained ability estimates by utilizing the information contained in each answer (correct or incorrect) to an item. Each item option is described by an item option characteristic curve.

Item response theory was introduced by Lord (1980). Item response theory models the relationship between a person's level on the trait being measured by a test and the person's response to a test item or question. Because the trait levels are unobservable, item response theory falls into the general class of latent trait models. In contrast to the classical test theory, item Response theory makes strong assumptions about an examinee's behaviour when answering the test items.

Dimensionality of the latent space, Local Independence and item characteristic curves are three important notions that arise in connection with latent trait models. Latent trait models that assume that a single latent ability is necessary to adequately account for an examinee's test performance are referred to as unidimensional. The method of factor analysis holds the most promise for developing and for testing a unidimensional test (Lumsden, 1961).

The assumption of local independence states that the probability of an examinee answering a test item correctly is not affected by his or her performance on any other item in the test. This means that for a given latent value, item scores are independent of one another. If the art of answering one item or the knowledge of one item answer influences the answering of other items, then the items are not locally independent (Allen and Wendy, 1979). Hambleton, Swaminathan, Cook, Eignor and Guilford (1978), define an item characteristic curve as a mathematical function that relates the probability of success on an item to the ability measured by the item set or test that contains it, that is: $P_{ig} = f(\theta)$.

An item characteristic curve is a graphical display of the relationship between the probability of passing a particular item and the examinee's position on the underlying trait measured by the test. Since scores on underlying trait are generally not observable, observed test scores are used as estimator of trait values. The two principal modes of determining the item characteristic curve are the normal Ogive and the logistic item characteristic curve (Baker, 1977)

Latent trait models are usually differentiated by the number of parameters estimated for the items and the nature of the item characteristic curve. The one parameter, also known as the Rasch, model (Rasch, 1966) is a special case of the item characteristic curve in which the parameter for discriminating power is assumed the same for all items and is absorbed in the unit of ability estimate (Douglas 1980). This means that in the one-parameter model, test items are described only in terms of their difficulties. Thus items which survived the one-parameter scaling discriminate equally (Bryce, 1981).

In its mathematical form, this model states that the observed answer, ani of person n to item I is governed by a binomial probability function of person ability, Zn and item easiness, Ei. The probability of a correct answer (Wright & Panchappkesan, 1969) is:

$Pr(a_{ni}=1) = Z_n E_i/(1+Z_nE_i)$, and the probability of incorrect answer is:

$Pr(a_{ni}=0) = 1-Pr(a_{ni}=1) = 1/(1+Z_nE_i)$

The model parameters Z_n and E_i are usually written in log form as follows:

$Pr(a_{ni} = 1) = \exp(b_n - d_i)/1+\exp(b_n-d_i)$

The one-parameter latent trait model provides a good fit for a typical achievement test data because it predicts the well-known S-shaped ogive for the item characteristic curve. An inspection of the above equation reveals several important features of latent trait models that can be realized in achievement test data that fit the model. First, the relationship between the response to an item probability and the parameters is non linear. The parameters are contained in the exponents. Second, person ability and item difficulty combined additively in the exponent. This implies that ability and item difficulty have been located on a common scale of measurement, namely position on the latent trait. Also implied by additivity is that person ability and item difficulty have equal weight in determining the effective response, potential, $Pr(a_{ni} = 1)$.

Where b_n = Log Z_n, and d_i = -Log E_i. The one-parameter model makes the following assumptions (Anderson, Kearny & Everrei, 1968; Tinsley & Dawis, 1975; Ficher 1976)

1. Items are scored dichotomously, that is, 0, 1
2. The total number of raw scores obtained on an item is a sufficient statistic for the estimation of the ability parameter.
3. The ability parameter can be estimated without bias by means of whatever section of items from the total test, provided the item fits the model. This means that the model ensures objectivity in measurement.
4. The estimation of the item parameters is independent of the distribution of ability parameter (θ) in the validating sample.
5. The validity of the model can be tested independent of the distribution of ability (θ) and the item difficulty (bi)

6. The probability of a correct answer to an item (i) by an individual (θ) is a function of the ratio, easiness-ratio.

Birnbaun (1968) proposed a latent trait model in which the item characteristic curve takes the form of a two-parameter logistic distribution. The two-parameter model describes items by both their difficulty and discrimination indices. The equation for the two-parameter model is expressed as:

$$P(x_{ij} = 1) = \frac{e^{D}a_i(\theta_j - b_i)}{1 + e^{D}a(\theta_j - b_i)}$$

Where X_{ij} is the answer to item I by person j, θj is the ability parameter for person j, bi is the difficulty parameter for item I, and e is the base of the system of natural logarithms, D = 1.7, a scaling factor and ai. Is the discriminating power. The three parameter model adds a guessing parameter as the third descriptor of the item characteristic curve. The three-parameter model is expressed as:

$$P(x_{ij} = 1) = ci + (1\text{-}ci)\frac{e^{D}a_i(\theta_j - b_i)}{1 + e^{D}a_i(\theta_j - b_i)}$$

Where c_i is the guessing parameter. Lord (1980) introduced the 5 - parameter model assuming parameters for Skewness and Kurtosis of the underlying distribution.

vi. **Item Analysis**

Ever since Binet and Simon(1916) plotted the proportion of correct answer to an item as a function of age, evaluation of the essential qualities of a test has been an important component in the field of educational measurement (Baker, 1977). Two basic theoretical models, the classical and the latent trait have been utilised and a wide variety of item analysis procedures have been developed using these approaches.

(a) Classical Item Analysis

The classical assessment of the qualities of the items in a test can assume both qualitative and quantitative dimensions. According to Anastasis (1976), qualitative item analysis addresses the content and form of the test. This involves the consideration of content validity and evaluation of item in terms of effective item writing procedures. According to this standpoint, test validity can be examined by relating the test to a defined domain (Pophan, 1971). Once an item is included in the domain, no empirical information or test analysis can or should lead to reformulating the item or eliminating it from the domain (Pophan, 1971; Anastasis, 1976).

This apriori approach to item analysis assumes that items are best developed using item domain specification. The problem with this qualitative analysis is that item writing always involves some subjectivity and that to eliminate the consequences of this subjectivity, analysis of empirical data is highly indispensable (Pophan, 1971). The second approach to this qualitative item analysis is a judgemental approach in which content specialists are retained. The judgemental task may assume a form consisting of expert directly rating the item objective congruence. Agreement among judges, experts, independently classifying the item is assumed sufficiently high to warrant the conclusion that raters tend to categorise best items with respect to the cognitive processes sought in the items (Pophan, 1971).

On the other hand, quantitative item analysis is concerned with the statistical properties of the item. Essentially, this entails the assessment of the item difficulty wide variety of item analysis procedures have been developed using these approaches.

(a) Classical Item Analysis

The classical assessment of the qualities of the items in a test can assume both qualitative and quantitative dimensions. According to Anastasis (1976), qualitative item analysis addresses the content and form of the test. This involves the consideration of content validity and evaluation of item in terms of effective item writing procedures. According to this standpoint, test validity can be examined by relating the test to a defined domain (Pophan, 1971). Once an item is included in the domain, no empirical information or test analysis can or should lead to reformulating the item or eliminating it from the domain (Pophan, 1971; Anastasis, 1976).

This apriori approach to item analysis assumes that items are best developed using item domain specification. The problem with this qualitative analysis is that item writing always involves some subjectivity and that to eliminate the consequences of this subjectivity, analysis of empirical data is highly indispensable (Pophan, 1971). The second approach to this qualitative item analysis is a judgemental approach in which content specialists are retained. The judgemental task may assume a form consisting of expert directly rating the item objective congruence. Agreement among judges, experts, independently classifying the item is assumed sufficiently high to warrant the conclusion that raters tend to categorise best items with respect to the cognitive processes sought in the items (Pophan, 1971).

On the other hand, quantitative item analysis is concerned with the statistical properties of the item. Essentially, this entails the assessment of the item difficulty (facility value) and item validity (discrimination, biserial and point biserial). These indices are organised into three broad categories representing the three classical approaches to item analysis: discrimination index; biserials correlation co-efficient; phi-coefficient and factor analysis (Henrysson, 1971).

When an item is scored either 1 or 0, the simplest index of its difficulty is its mean item score, p. This value is also the empirical probability that the particular examinee involved will pass the item (Guiford, 1954). While the statistic, P can be used to rank or order items for difficulty; it is not the best index for item analsysis. This is because item difficulty suffers like all percentages from the limitation that the index cannot be regarded as estimated in equal intervals (Wood, 1990). The difference (intensity of difficulty) between items with facilities of 0.40 and 0.50 is not the same as the difference between the items with facilities 0.10 and 0.20.

This "elasticity" in the unit makes for misjudgement in comparing facilities and rules out simple statistical manipulations of the difficulty index, P, except for summation which results in the estimate of the mean total test (Wood, 1990). Furthermore, item difficulty is inversely related to difficulty and is, therefore, a direct measure of easiness. In addition, item difficulty does not provide a linear unit of measurement that is, it is not linearly related to ability (Herman, 1990).

A discrimination index is meant to communicate the power of an item in separating the more from the less capable on some measuring attributes. The simplest discrimination index is the mean difference Pu-Pl, where Pu and Pl are the proportions of passing examinees in the upper and lower groups respectively (Wood, 1990). Lawshe (1942) in considering the labour involved in estimating item discrimination indices for test of several items developed a monograph for estimating the validity of the test items, D-values.

Much work on item discrimination power has been done using the top 27 percent and the bottom 27 percent because this division proves the greatest efficiency in estimating item discrimination (Kelley, 1937). In the upper-lower 27 percent method, the number choosing the different answer alternatives are compared for the 27 percent of students with the highest criterion scores and 27 percent with the lowest and dividing by one - half of the total number of the student.

The discrimination index is operationally computed as a correlation either biserial or point biserial between success and failure on an item and the total score. The idea is that the higher the correlation between examinees scores on an item and their scores on the test, the more effective the item is in separating them (Wood, 1990). The biserial correlation co-efficient gives an estimate of the Pearson-product moment correlation between the criterion score and the dichotomised item.

The biserial correlation is based on the assumption that both variables are continuous, but that one has been divided at one point into two groups, those who passed the item and those who failed it (Wood, 1990). The two variables are each assumed to follow a normal distribution in the population of examinees from whom the sample is drawn. The principles upon which the biserial co-efficient is based is that with Zero correlation there would be no difference between the means of the passing and failing groups, and the larger the difference between the means the larger the correlation (Wood, 1990).

The computation of biserial is given by the expression (Kuang, 1952; Downie & Heath, 1974):

$$\text{Biserial (rbis)} = \frac{\overline{X}p\ \overline{X}t}{SD_t} \times \frac{P}{y}$$

Where $\overline{X}p$ represents the mean score of all students answering the item correctly, $\overline{X}t$ is the mean score for the entire group and SD_t is the standard deviation of the distribution of the criterion scores. P is the proportion of the total group answering the items correctly, and y is the ordinate of the normal probability table cutting off above it an area equal to percentages of passing groups.

The point-biserial (rp^b) is also used as an item discrimination index, where one of the variables (total scores) is regarded as continuous and the other (item score) can take on 1 for correct and Zero for incorrect responses (Wood, 1990). The point biserial has the formula (Downie & Heath 1974):

$$rp^b = \frac{M_R - M_T}{SD_T} \times \sqrt{\frac{p}{q}}$$

Where M_R is the mean of the passing group, M_T is the mean of the total group and SDt is the standard deviation of the entire group. P is the proportion of the total group answering the item correctly, and $q = 1 - P$.

The point-biserial (rp^b) is lower than the biserial (rb). It has been shown (Wood, 1990) that the point-biserial (rpb) is equal to 0.82rb. The point-biserial correlation is influenced by the item difficulty and makes more contribution to the functioning of the test than the biserial correlation (Wood, 1990). Furthermore, (Wood, 1990) argues that point-biserial is effectively a combined index of difficulty and discrimination. Because it is less influenced by item difficulty, the biserial correlation is relatively stable from one examinee's group to another (Lewis, 1974). The biserial plays a part in item characteristic curve theory, in the equation for estimating the steepness parameter of the curve. The point-biserial does not (Henrysson, 1971). Again, the point-biserial fits directly into the algebra of multivariate analysis, the biserial does not (Lewis, 1974). When it comes to estimating test parameters from item parameters, it is the point biserial which is needed (Wood, 1990).

Another argument put forth in favour of the point-biserial (pbr) is that it tells more about the contribution from the particular item to the predictive validity of the test than the biserial. According to Wood (1990), this is explained by the fact that the point biserial favour items of average difficulty, that is, items with facility value around 0.5 than does the biserial (rb). This means that the point-biserial correlation co-efficient is a combined measure of item-criterion relationship and the difficulty levels.

A rank biserial correlation co-efficient was suggested by Glass (1996) for situations where the criterion scores are reported in ranks in a group. The correlation co-efficient proposed by Glass has the formula:

$$\text{Rank biserial (Pb rank)} = \frac{2}{n}(V_R - V_W)i$$

Where, n = the number of subjects

VR = the average rank of those scoring right

Vw = the average rank of those scoring wrong.

This co-efficient gives an estimate of the correlation between the two series of ranks as expressed by Spearman rho (Glass, 1996). Employing the rank biserial correlation, the dichotomous score is regarded as having ranking variables underlying it (Glass, 1966). In this regard, it is analogous to the assumption made In biserial correlation, that is, a normally distributed variable has been forced into a dichotomy. The phi-co efficient is a measure of the relationship between an item and the criterion. It is estimated from a four-fold table based on the proportion of cases passing and failing each item in the upper and lower groups. Its value ranges from –1.00 to 1.00. This co-efficient assumes a genuine dichotomy condition under which it was obtained (Neil & Jackson, 1970).

One major disadvantage in the use of product moment correlation co-efficient as an index of item validity is that the item indices do not Remain constant over the entire test (Lord and Novick, 1968; Wood, 1988). In consideration of this limitation in the use of correlation techniques for item validation, Finney (1944) suggests the application of the theory of probability to the problem of item analysis. According to Kuang, (1952), the first step in the probit analysis is to divide the examinees into ability categories of equal interval and convert each proportion into standard scores, Z plus 5. The addition of the constant 5 is to get rid of negative numbers. The value z plus 5 is called a probit.

The next step is to plot probits as a function of the criterion groups. The process of estimating the unknown parameters of the limen is to find the best fitting linear relationship between ability and the probit of the percentage passing the item by the method of maximum likelihood (Kuang 1952). Kuang defines the limen as the point measured in standard deviation unit of ability where the probability of an examine of that ability either passing or failing the item is 0.5.

In view of this principle, Kuang suggests that the slope of the limen line may be defined as the discrimination index. The principle underlying item analysis by probit technique is to divide the sample of examinees tested into ability groups and to compare percentage of successes in prohibit value of an item in the various groups. Another classical technique employed to assess the quality of a test item is factor analysis. Factor analysis is a statistical technique which has been widely used as a further step towards analysis of the test item. Factor analysis is a multivariate statistical technique which is used for analysis relationships between correlation co-efficient in a set of data, identifying psychological constructs, data reduction, confirmatory and exploratory analysis (Joe, 1992; Pointerotto, Burkard & Riegard, 1995).

Three different approaches to the problem of obtaining estimates of the factor loading have been identified: maximum likelihood; principal components, and the centroid methods (Browne, 1968). Lawley (1943) used the method of maximum likelihood and Rao (1956) used a method based on the theory of canonical correlations to derive estimates of factor loading. The principal components method was originally developed by Hotelling and applied to the correlation matrix with diagonal elements replaced by the communality estimates. In this solution, squared multiple correlations are used as initial approximations to the communalities (Browne, 1968). Before the advent of computers, factor analytic studies were carried out using Thurstone's centroid method which yields an approximation for its communality (Browne 1968). There are basically three steps in any factor analysis (Joe,1992; Lemk & Wiersma, 1976):

1. Obtaining correlation co-efficient matrix for a number of scores. This matrix contains the indices of correlation of every measure with every other measure. It also contains indices selected to represent the correlation of each measure with itself (diagonal values). The correlation matrix may be that between variables (R-Type analysis) or between objects, individuals (Q-Type analysis). This study employed the R-type analysis.
2. Once the correlation matrix has been completed, the process of factor extraction, initial factoring is undertaken. There are principally two ways of initial factoring, the principal components method and the common factor method. The result of this process is a second tabular display, called a table of factor loadings. This table shows the number of factors needed to account

for the values in the matrix. The loadings are correlation co-efficient between tests and factors.

3. Rotating the number of variables obtained from initial factoring to arrive at a final solution. Factors are rotated in order to determine the underlying structure behind them. Rotated factors can be orthogonal or oblique (Guilford, 1954). In orthogonal rotations the factors are independent of themselves and the axes maintained their 90 degree separations after rotation as before. Orthogonal solutions include varimax and quartimax rotations. The varimax rotation sample each variable in the factor structure. In oblique rotations, the axis are not required to maintain their 90 degree separations, each one can be rotated through its own angle. Oblique rotation includes promax and oblimax solutions (Joe, 1992).

The loadings in the factor matrix can be squared to obtain the proportion of predictable variables in the test that can be accounted for by that factor (Lemk & Wiersma, 1976). There are basically three types of factors: general factor, bipolar factor and group factors. A general factor has moderate to high loadings for all tests. A bipolar factor has both high positive and high negative loadings, while the group factor has high loadings but at least one test has a near zero loading. To determine the proportion of variance due to each factor, the squared loadings are summed for each factor. The sum of the squared loading is called the eigen value. It is the sum of the squared loadings of the variables on each factor To obtain the proportion of variance due to a particular factor, the eigen value is divided by the number of tests(Variables) and multiplied by 100. Lemk and Wiersma (1976) defined communality, h2 as the proportion of the variance of a test due to all factors. To compute the communality, the rows of the table are summed across. The sum of the eigen values is equal to the sum of the communality (h^2).

b) Latent Trait Item Analysis

Practical application of the latent trait model requires some methods of estimating the item parameters of the model. The estimation procedures in latent trait models range from adhoc procedures to maximum likelihood procedures. Lord (1968); Urry (1974) and Jensena (1976) employ adhoc approaches to item analysis in the latent trait model. The first approach is the correlation method, while the second approach is the regression method, item characteristic curve.

Lord (1968) employs the correlation approach for simultaneous estimation of the item discrimination indices for all items in a test. To accomplish this he factor analyses the matrix of inter-item correlation. In this solution, the parameter estimating item difficulty is evaluated by the normal deviate corresponding to the proportion of subjects in the total group answering the item correctly, while the discrimination power was estimated by the loading of the item in a one-factor analysis. This factor analytical approach is termed "heuristic" approach (Lord, 1968; Henryson, 1971). Jensema (1976) proposes a direct conversion method for estimating initial item parameters for the latent trait model from the item parameters of the classical test model. Urry (1974) and Jensema (1976) show that latent trait discrimination index (ag) for an item g is estimated from the classical test point biserial by the expression:

$$ag = \frac{p\theta}{\sqrt{1 - p_2\,(\theta)}}$$

$P(\emptyset)$ = correlation between an item and the latent ability and represents the factor loading of the item obtained from the matrix of tetrachoric correlation.

The regression method involves the regression of an item on the latent trait and is based on the concept of item characteristic curves (Baker, 1977; Drasgow and Lissak, 1983). Here the discrimination power is the steepness of the item characteristics curve. The item difficulty is the value of the ability (σ) at which the probability of a correct answer to an item is 50 percent. If we assume a normal distribution of ability, then b, item difficulty can be estimated from the proportion passing an item. If this proportion is designated dg, for difficulty of item g, then the latent trait item difficulty is estimated as (Ross, 1966)

$$bg = N^{-1}(dg)\sqrt{2.98 + a^2_g}$$

Where $N\text{-}1(^d g)$ is the normal deviate corresponding to the proportion passing the item g; ag is the discriminating index for item g, and 2.89 is a scaling factor, a constant.

Although the factor analytic approach to item parameters estimation in the latent trait model is relatively simple, some workers such as Lord (1968); Wright & Panchapakesan (1969); Bock and Lieberma (1970) have preferred to employ estimation procedures from the field of statistics, primarily maximum likelihood estimation. Within this framework, Bock & Liberman (1970) distinguished between unconditional and conditional maximum likelihood. Under the conditional approach, the item parameters are estimated conditional upon the subject ability scores.

In this approach, the basic data for item parameter estimation are a proportion of correct answers (difficulties indices). Under the unconditional approach, the role of ability is removed in the estimation equations, hence the item parameters are estimated unconditionally with respect to the latent ability dimension (Wright and Panchapakesan, 1969; Bock & Lieberman, 1970; Wainer, Morgan & Gustafsson, 1980). In this approach the basic data used for item analysis are the patterns of a subject's score, one(1) if correct and zero (o) otherwise.

The pattern of answers is then used in the maximum likelihood equations to obtain the likelihood function or the joint probability distribution (Choppin, 1980; Soriyan, 1972). The unconditional approach to item parameter estimation involves the following steps (Bock & Lieberman, 1970):

1. The determination of the likelihood function of the sample scores.
2. The ability function is integrated with respect to the likelihood function in order to obtain an expression for the likelihood of the item parameters.
3. Indices of the item parameters are estimated by the principle of maximum likelihood, and
4. These estimated values of the item parameters are then used to estimate the ability of a given examinee in terms of his scores in the validating sample (See Appendix viii)

In this approach, the role of ability is removed from the likelihood equations, hence the item parameters are estimated unconditionally with respect to the latent ability. Under the conditional approach, the likelihood function is expressed as a conditional function in terms of only the item parameters (Lord 1968; Anderson, 1973; Wright & Douglas, 1977). The conditional maximum likelihood parameter estimates were obtained by solving the resulting conditional equations simultaneously with ability estimates by an interactive technique. This estimation procedure was termed conditional in that abilities of the subjects are as unknown (Baker ,1977).

Wright & Douglas (1977); Izard & White (1980) developed an approximation procedure (PROX) for item analysis based on the latent trait model. PROX analysis requires answers given by students to be listed in a student-by-item matrix. The calculations are carried out on marginal totals of correct and incorrect frequency (count) for both items and subjects. The procedure eliminates any subject with perfect or zero scores. Items not attempted by any students are also deleted from the matrix before the calculation.

This method facilitates item validation as it can be handled using a hand calculator. Using the PROX procedures, the item difficulty (bi), for item I is

$$bi = \ln \left[\frac{N - s}{s}\right]$$ times item difficulty expansion factor due to sample spread.

N = Number of examinees

S = Total scores on an item

ln = The base of the natural logarithms

Test analysis techniques based on latent trait model attempt a separation of the respective contributions of person, ability (ø) and item parameters. Utilising PROX procedure, Izard & White (1980) show that the ability estimate is given by:

ability (ø) = ln [r/n-r] times person ability expansion factor due to test width.

Y = Subject total score on the test

n = number of items composed of the test

(See Appendix vii).

The development of a valid and reliable test requires consideration of quantitative information regarding the difficulty and item validity of each test item. Two methods of item analysis, classical and latent trait models were reviewed. Within the classical test model, four statistics biserial and point biserial, phi-co-efficient, probit analysis and factor analysis were reviewed as indices of item validity. Recently, there has been a marked shift in item analysis techniques from the correlationally based psychometric approach to those estimation procedures used by statisticians, in particular, the use of maximum likelihood procedures for the estimation of item parameters. The most significant concept to arise out of the application of maximum likelihood item analysis procedure is the distinction between unconditional and conditional estimation made by Bock & Lieberman (1970). As latent trait theory becomes complete, the role of ability estimates will become more prominent. This study validated a physics Achievement Test using PROX approach as exemplified in the work of Izard & White (1980).

v. Connection of the Latent Trait item Parameter with the Classical Item statistics.

Person, ability and item parameters of the latent trait models do not correspond directly to the indices of item difficulty and item validity of the classical test model. Nevertheless, the latent trait item difficulty, person ability, and discrimination power are closely related to the classical p-values and item validity. The latent trait models provide valuable insight into the classical concerns for test reliability and validity.

The classical approach to item difficulties uses proportion of persons attempting the item who are successful. Latent trait item difficulties can be approximated from the classical item p-values, using the formula (Wright, 1990):

D = M + Y log (I-P)/P

where D = Latent trait item difficulty

M = Mean score on the entire test

Y = $(I + S^2/1.7)1/2$

S = Standard deviation of the score, and

P = Classical item p- value

The values D, M and S are in logit and the constant, 1.7 transforms the normal distribution to follow the logistic. The formula removes the short comings of the classical p-values by first transforming them into a linear scale by the logit function, log(1-P/P) and this transformed P-value is then reconverted so that the influence of the sample standard deviation, s, and sample mean, M, are removed. If abilities are normally distributed, the resulting item difficulty, D, is invariant (Wright, 1990). This means that difficulties of items can be compared even though they might come from different sample of subjects.

The classical test approach to reporting a person ability is to count the number of correct answers obtained and to report either this raw score or on some norm-based transformation of it. Under the assumption that items are normally distributed, latent trait ability estimates can be approximated from raw scores with the formula (Wright 1990);

B = H + X log (r/l-r)

Where X = 1 + V/2.891/2

B = subject's ability estimate

H = mean difficulty of the test

V= the variance of these difficulty indices

γ = subject's raw score, and

L = number of item composed of the test.

The disadvantages of the raw scores are removed in two steps: First by transforming the scores into an interval scale using the transformation: log (r/L-r) and second removing the influence of the mean and standard deviation of the test item difficulties.

Ross (1966); Jensema (1976); Wood (1990) suggest some connections of biserial correlation co-efficient with the latent trait item parameters. If the underlying trait ability (ø) has a normal distribution, then the discrimination index (ag) of the latent trait parameter bears a simple relation to the biserial correlation co efficient (Wood, 1990). Wood expressed the relation as:

$$ag = \frac{r \text{ biserial}}{\sqrt{1 - r^2 \textit{ biserial}}}$$

Furthermore, in a population in which the underlying trait is normally distributed Ross (1966) suggests a simple relation between the latent trait item difficulty (bg) and the biserial correlation. The relation is expressed as:

$b_{(g)} = Y_{(g)}$ /r biserial

Where Y_g = normal deviate corresponding to the proportion of success on item (g) in the total sample.

vi. Reliability and Standard Error of Measurement.

The reliability of the test is intended to specify the degree of accuracy, dependability or consistency with which the test measures the variable it is designed to measure (Thorndike, 1990). The concept of reliability has been under continuing reformation and redevelopment with the resulting increase in clarity and range of applicability. Cronbach, Rajaratnam & Gleser (1969) have summarised this development within the framework of the classical test model, which assumes that the observed test score for each examinee can be resolved into two components; an unknown true score and a random error.

There are several ways of defining and interpreting reliability. The following interpretations have been summarised by Lord & Novick (1968); Allen and Wendy (1979). Interpretation one states that the reliability of a test equals the correlation of its observed scores with observed scores on a parallel test. Interpretation two states that the reliability co-efficient is the ratio of the true score variance to observed score variance; that is, reliability $=V_{arT}/V_{arX}$. Interpretation three presents the reliability co-efficient as the square of the correlation between observed and true scores; that is, reliability = $r^2(X,T)$. Interpretation four states that the reliability co-efficient is one minus the squared correlation between observed and error score. That is, reliability = 1-r2 Interpretation five relates reliability to error score variance and observed score variance. That is, reliability is one minus the ratio of error score variance and observed score variance (1- Var E/ Var. x) Lumsden (1961) advocates three reasons for estimating reliability co-efficient as:

1. to guide test selection
2. to support inference about test score standard error of measurement and
3. to support inference about the validity of a perfectly reliable test.

Methods for computing reliability of a test are of four types. The test / re-test reliability yields a reliability estimate that is based on testing the same individual twice with the same test and then correlating the results. Stanley (1971) explicates some complications inherent in this method. The most serious problem with the test/retest is the actual memory of particular items and of previous response to them. This carryover effect leads to contamination of scores. To circumvent this problem, Allen & Wendy (1979) suggest a re-arrangement of the order of the item and or the test option of each item each time the test is administered. Another problem with the test/retest method is that uncooperative examinees might object to the second testing and they may, therefore, give haphazard responses. This will operates to lower the correlation between the two testing (Guilford, 1954). Test/ retest gives a measure of stability co-efficient.

A parallel test forms reliability estimate is the correlation between observed scores on two parallel test. The requirements for parallel or equivalence forms test are that the two forms have the same common factors and the total variance equal (Livingston, 1988). The problem with this method is that in preparing parallel forms, there is the danger that the two forms will vary so much in content and format that each will have substantial specific variance distinct from the other. In which case the reliability will be underestimated (Stanley, 1971). There is the reverse danger that the two forms may overlap in specific details, in content, such that variance due to specific sampling of content may be common to the test. In such a situation, the obtained correlation will over-estimate the independent form reliability. Parallel test administration given virtually at the same time gives a measure of co-efficient of equivalency. The co-efficient using comparable forms test with an interval between testing is a co-efficient of equivalency and stability.

Reliability via subdivision of a single total test, called the internal analysis approach. A test is said to be internally consistent if all of its items measured the same things (Martuza, 1971). Internal consistency reliability is estimated using only one test administration and thus avoids the problem associated with repeated testing (Stanley, 1971). The most widely known method using this approach yields a split-half reliability estimate. One of such procedures subdivides the total test artificially into two half length tests and correlates with the scores. This coefficient of correlation estimates the reliability, not of full test, but a test only half as long (Martuza, 1971).

To adjust the value of this reliability, the Spearman-Brown prophecy formula has been suggested. A simplified version of which is.

$$\text{Reliability} = \frac{2 \times \text{reliability of half test}}{1+ \text{reliability of half test}}$$

The main limitations of this method (Rastogi, 1983) are:

1. It is not always possible to split the test into two halves. This is moreso, of tests of skills, speed.
2. since the reliability of the test depends upon the scores obtained by a single administration, it is more likely to be affected by chance error than the test/ retest reliability.

Another way to compute the split-half reliability co-efficient is by using the Guttman formula (Guttman,1945). In this formula, if V_1 and V_2 represent the variances of scores on the two half tests, and v total represents the variance of scores on the full test, then the split-half reliability of the whole test is given as:

$$\text{split-half reliability} = 2\frac{(1 - v_1 + V_2)}{\text{v total}}$$

Two methods for forming test halves are commonly suggested (Allen & Wendy, 1979). The first method called the odd/even methods classified items by whether they are odd or even numbers on the test. A second method is to form the halves in order. Each examine obtains a score on the first half of the test and a score on the second half of the test.

The other method of computing reliability is via utilising the variances and intercorrelations of test items. The second groups of internal consistency procedure is based essentially on the analysis of variances among individual items and was originally devised by Kuder & Richardson (1937), the most widely used version formula 20 is given:

$$\text{Reliability} = \frac{n}{n-1}\left[\frac{\sum p_i q_i}{1-v}\right]$$

Here n = number of items in the test

P1 = the proportion of examinees answering correctly on items i

q1 = the proportion of examinees answering wrongly on item I and

v= total test variance.

A lower estimate of this value is provided by Kuder-Richardson formula 21, which takes the form:

$$\text{Reliability} = \frac{n}{n-1}\left[1 - \frac{(\bar{x})^{2/n}}{v}\right]$$

where x = total score
$\bar{x}$ = mean of the test score

The most general form of the analysis of item variances is provided by

Cronbach's co-efficient alpha (1951), the formula which is expressed as:

$$\text{Reliability} = \frac{n}{n-1}\left[1 - \frac{\sum V_i}{\sum V_{total}}\right]$$

where n= number of items in the test

V_i = variances for item I

V_{total} = variance of the total test.

Common to all the reliability estimation procedures is the sample based nature of all the estimation procedures (Weiss & Davision, 1981). Reliability estimates are specifically a function of a particular set of items and sample of subjects on which the data were collected. This is a major weakness of the classical test model.

The accuracy of any item or ability measurement is an important consideration in the development, application and interpretation of a test (Wright & PanchaPakesan, 1969). One of the main advantages of the latent trait model is that the concept of reliability, is not used. In place of reliability, latent trait models use the concept of standard error or precision of the measurement (Wood, 1990, Korashy, 1975). Birbaum (1968) defines stand error as the measurement effectiveness of a test item at each level of the trait, ability being measured.

In latent trait models, each item validity or person measure, ability estimate is accompanied by its standard error, the most rudimentary indicator of reliability in the classical test model (Korashy, 1995). Unlike the classical reliability estimates, the standard error of measurement is independent of the particular examinee sample and it is an indication of the amount of error in ability estimate at different points of the ability continuum (Lord & Norvick, 1968). Each item contributes to the standard error independently of all other items in the test.

In the classical test model, the contribution of each item to the test reliability and validity depends upon what other items are in the test (Wood, 1990). The standard error associated with a test or person measurement (ability estimate) is evaluated using the PROX Procedures (Izard & White, 1980). The standard error associated with an item is given by:

$$\text{Standard error} = B \times \sqrt{\frac{N}{S(N-S)}}$$

Where B is the item difficulty expansion factor due to sample spread

N= number of subjects, and

S= score on an item

The corresponding value associated with ability estimate is given by:

$$\text{standard error (ability estimate)} = A\sqrt{\frac{n}{r^{(n-r)}}}$$

Where A = person ability expansion due to test width

n = Number of items

r = person's score on the test (see Appendix vii).

Another measure of standard error is the RASCH STANDARD ERROR (Weiner & Wright,1980). Rasch standard error is given by:

$$\text{standard error} = \sum[p_{ij}(1 - p_{ij})]^{-1/2}$$

Where p_{ij} is the probability of a subject, answering item I correctly. For an efficient precision of measurement the value of the standard error must not exceed $\sqrt{\frac{1}{f}}$, where f is the number of ability groups (Bryce, 1981). Reliability is defined as: the proportion of observed score variance divided by the true score variance, the squared correlation between true scores and observed scores, the correlation between observed scores on two parallel tests. Several estimates of reliability co-efficient are widely used including test retest, parallel forms and consistency estimates. In contrast to the classical test reliability, latent trait psychomepic models employ the notion of standard error of measurement. The standard error is interpreted as a measure of the precision of the estimate of the examinee's ability on the test items. The standard error associated with an item is not influenced by the characteristics of the other items in the test. This is in contrast to the classical test model where the contribution of one item to the test reliability is not easily identified, (Keats 1980), hence the usefulness of latent trait models is the development and standardisation of a physics achievement test.

(vii) Validity and Analysis of fit.

The question of test validity concerns what the test measures and how well it does so. A test has validity or is valid if it measures what it purports to measure. Shontz. (1965:235) gives the following opinions about validity.

> To be valid is to be bonafide, genuine, authentic, truthful and effective. Validity of measurement is the extent to which tasks, observational systems, questionnaires and ratings are effective, truthful and genuine in serving their stated purpose. Validity of measurement is the soundness of the interpretations of a test or observational system. The one characteristic which a measurement procedure must have is validity.

Validity can be assessed in several ways. The American Psychological Association (APA) committee recommendations (Anastasi, 1976) classified these procedures under four categories designated as content, predictive, concurrent and construct validity. Content validity involves essentially the systematic examinations of the test content to determine whether it covers a representative sample of the behaviour domain to be measured (Anastasi, 1976). Such a validation procedure is commonly used in evaluating achievement tests which are designed to measure how well the individual has mastered specific skills or courses of study.

Content validity is accomplished by a mere inspection of the content of the test to ensure its relevance and comprehensiveness (Shontz, 1965). The soundness of descriptive interpretation asks such questions as:

> Do the observations truly sample the universe of tasks the developer intended to measure or the universe of situations in which he would like to observe? Does the test measure an important educational out come? Does the battery of measure neglect to observe any important details (Cronbach, 1971:44).

Content validity is not synonymous with face validity. Face validity refers not to what test actually measures, but to what extent it appears superficially to measure (Anastasi, 1976). That is, face validity pertains to whether the test looks valid to subjects who take it and the administrative personnel who decide upon its use. Predictive validity indicates the effectiveness of a test in predicting some future outcomes usually referred to as the criterion. The criteria most frequently cited in test manuals include: general academic achievement, performance in specialised training and on the job performance (Stanley, 1971). The relations between test scores and indices of criterion status obtained at approximately the same time are known as concurrent validity (Allen & Wendy, 1979). Anastasi, (1976) contrasts predictive with concurrent validity. She is of the opinion that the logical distinction between predictive and concurrent validity is based not on time, but on the objectives of testing. Sax (1989) refers to construct validity of a test as the extent to which the test may be said to measure a theoretical or a psychological construct.

Construct validity is the extent to which the test measures the attribute or phychological trait it is said to measure. Procedures for construct validation include: age differentiation, correlation with other test scores and factor analysis (Anastasia, 1976). Multi trait- multi method validity is an aspect of construct validity that was developed by Campbell & Fiske (1959). This method is used when two or more traits are being measured by two or more methods. There are two main multitrait-multimethod validity: convergent validity and discriminate validity. Convergent validity is demonstrated by high correlation between scores on tests measuring the same trait by different methods. Discriminate validity is demonstrated by low intercorrelations between scores on tests measuring different traits (Allen & Wendy, 1979). Another form of construct validity is the factorial validity which is validated by means of factorial analysis.

The usual statistics employed to assess the internal validity of a test in the classical test model are the item biserial (Youngman, 1979:Wood, 1990). Since the magnitude of this item statistic depends on the ability distribution of the sample, it has the disadvantage of being sample dependent (Douglas, 1990). Within the latent trait test model, the internal validity of a test is assessed in terms of the statistical fit of each item to the model. The analysis of fit is a check on internal validity. If the fit statistic of an item is acceptable, then the item is valid (Korashy, 1995) Gallini, 1983; Wainer, Morgan & Gustfson, 1980).

If a given set of items fit the model, this suggests that they refer to a unidimensional ability. Fit to the model also implies that the item discriminations are uniform and substantial, that there are no errors in item scoring. Fit to the model also implies that guessing has had a negligible effect on test scores (Bock & Lieberman, 1970; Wright & Panchapakesan, 1969). A measure of fit statistics commonly employed is the goodness of-fit (Anderson, 1973). This test involves a division of the sample into ability subgroups by score levels.

The proportion of subjects expected to answer an item correctly is compared to the actual proportion answering the item correctly. The difference between observed and expected scores are summed over groups, using the number of mean item difficulty (Gallini, 1983). Dixton & Massey (1969) suggested the criterion likelihood ratio test as a measure of fit to the model. The criterion likelihood ratio statistics, x2L is given as:

$$x2L = -2 \times \text{observed score} \times \log \frac{(\text{observed score})}{\text{Expected Score}}$$

Wright & Panchakesan(1969) introduced a test procedure, statistic y as a further measure of fit. This test procedure entails a comparison between expected and observed frequencies for all items in the ability groups and divided by the test variance, that is:

$$Y = \frac{nr_i - Enr_i}{(\text{variance})^{1/2}}$$

nri = observed score for item I

Enri = Expected score for item I

The expected score, $Enri = \frac{\exp(b - d)}{1+\exp(b-d)}$

b = ability estimate for each group

d = difficulty estimate for each group

$$\text{Fit statistic } Y = \sum_{r-j}^{k} y^2$$

Finally, a measure of fit to the model commonly used is the mean square residual (Smith, 1996). The between fit approach is based on item difficulties obtained from the ability groups. This statistic, $x^2(UB)I$ is based on subgroup

residuals and is given by: $X^2(UB)_i = \sum_{J-1}^{J} \frac{(x_{ni} - P_{ni})^2}{Nj}$

Where j is the number of ability groups, Nj is the number of subjects in each ability group, Xnj is the proportion of subjects responding correctly to item, I, given the over-all ability estimates for the subject and the difficulty of the item. This chi-square statistic is converted to a mean square by the degree of freedom J-1.

That is, mean square residual $MS(UB)j = \left(\frac{1}{j-1}\right)x^2(UB)_i$

The standard deviation, S of the mean square residual,

$SMs\ (UB)I = \left(\frac{2}{j-1}\right)^{1/2}$

The mean square residual is converted to a t statistics using the cube

transformation (Smith, 1996):$T = [(MS(UBI)\ 1/3 - 1)((\frac{3}{S})] + \frac{S}{3}$

where S is the standard deviation of the mean square.

An item will fit the model if it has a fit statistic of 1.50 and below (Bryce, 1981). A large positive fit statistic indicates misfit, no fitting. A low fit statistic nearer to one indicates better fit (Bryce,1981). A measure of fit to the model used in this study was the goodness of fit as exemplified in the work of Gallini (9830) using a physics achievement test.

In latent trait models, a fit to the model implies validity, that item discriminations are uniform and substantial, that there is no error in item scoring. Thus the criterion of fit to the model enables the test developer to identify and delete bad items.

2. Empirical studies Associated with Instrumentation Research

In order to establish empirical background of this study, studies utilizing the classical test model and latent trait models were reviewed.

Studies on Instrumentation Research Based on the Classical Test Model

Within the classical test model, a lot of indigenous studies have been undertaken on instrumentation research. This is exemplified in the work of Ogomaka (1984); Agwagah (1985); Nwankwo (1985); Nworgu (1985); Obioma (1985); Ndukwe (1990); Ogochukwu (1990).

Obioma (1985) developed and validated a diagnostic math's achievement test (DAMAT), using a 60- item, test on samples of students in Anambra, Benue, Cross River, Imo and Rivers States. On analysis of the results, the internal consistencies of the DAMAT and its subsets ranged from 0.83 to 0.77. The DAMAT exhibited low inter-correlation among its subsets. Furthermore, factorial validity was also established for the instrument.

Nworgu (1985) developed and preliminarily validated a physics Achievement Test (PAT) using a stratified sample of 564 students randomly drawn from a sample of 16 secondary schools in two education zones in Anambra state. The inter-correlation among the subsets of PAT were all positive, ranging from low to moderate correlations. Both the reliability and validity of the PAT were high. The unit reliability was found to be 0.06, those of the subsets ranged from 0.05 to 0.08. Agwagah (1985) developed and validated the achievement test in mathematics method (ATIMM). A 50 item ATIMM was developed and validated with a sample of 506 teacher training college students. The validity (content) was established by experts while a reliability co-efficient of 0.96 was estimated.

Nwankwo (1985) developed and validated an instrument for measuring Test Anxiety scale among secondary school students. The area of the study was all Education Zones of old Imo State which included: Aba, Okigwe; Orlu, Owerri and Umuahia with a sample of 2351 students using a 30-item Test Anxiety scale (TAS). The reliability estimates, stability and internal consistency were found to be 0.92 and 0.94 respectively. The concurrent and construct validity co-efficients of 0.96 and 0.90 respectively were established. The author also found that test anxiety increases with increase in class level.

Ndukwe (1990) developed and preliminarily validated an Integrated science Achievement Test (ISAT) with a sample of 1848 students from 42 secondary schools in five education zones of Imo State. The main data for the study was a 60 – item test based on four levels of cognitive domain. He reported a high positive correlation between the subsets of the integrated science test. The factorial validity of the scale was also established.

Nworgu and Habor-Peters (1990) developed and preliminarily validated a physics achievement test (PAT) based on the Senior Secondary School physics curriculum. The intercorrelation co-efficients among the content areas of PAT were high and positive. They reported a reliability co-efficient (KR-20) of 0.92. They concluded that the PAT was valid and reliable.

Mckenzie and Padilla (1986) constructed and validated a test of graphing in science (TOGs) with a sample of 410 students using a 20-item test of graphing in science. The reliability co-efficient (KR-20) was 0.83. Item discrimination and difficulty indices fell within the acceptable range for a reliable test. They concluded that the test of graphing in science was valid and reliable. As further evidence of test validity, the items were submitted to a principal factor analysis. The result showed that only one factor, accounting for 27.1 % of the variance had an eigen value above 1.0. Correlation with the one factor varied from 0.23 to 0.74. It was concluded that the test of graphing in science was valid and reliable.

Ogomaka (1984) constructed an 82-item students' Evaluation of Teachers Effectiveness scale (SETES) and validated with a sample of 800 class four secondary school students from old Imo State. Both the face and content validity of SETES was established. The reliability of the scale was found to be 0.89.

Ogochukwu (1990) sees the need to develop and validate an instrument (shyness scale) for the identification of shyness. The reliability (split-half) co-efficient of the instrument using the Spearman-Brown formula was 0.86. The discrimination capacity of the scale (point biserial) ranged from 0.34 to 0.68 and the inter-correlation co-efficient of the three sections of the scale ranged from 0.866 to 0.954. Indigenous workers have undertaken studies on instrumentation research based on the classical test model. Because of the numerous shortcomings of the classical test model, the present study is an instrumentation research aimed at developing and establishing the psychometric properties of a physics achievement test (PAT) for senior secondary physics students using the one parameter and the two-parameter latent trait models.

ii. Studies on Instrumentation Research using Latent Trait Models.

Much work has been done on latent trait test models but it is mostly foreign. Little work appears to have been done using latent trait model in achievement testing in Nigeria, as attested to the paucity of published work.

Douglas, Khavari and Farber (1979) in a study compared the classical item analysis with the latent trait model, the one parameter model. The sample was 373 predominantly whites university undergraduate located in an urban industrial city which is known to have a high rate of alcoholic beverage consumption in the united states. The subjects were administered the Mac Andrew alcoholic instrument and the Khavari alcoholic test.

The MacAndrew alcoholism test was a 49-item instrument designed to indirectly diagnose alcoholism, while the 12 items Khavari alcoholic test was an instrument for assessing quantity (frequency) of alcoholic consumption. The 49 MacAndrew items were anchored on five response alternatives: false, partly false, uncertain; partly true and true. For purpose of analysis, the five responses were collapsed into two categories. For example, responses of true and partly true were both treated as alcoholic response for the question: "I have not lived the right kind of life". The Khavari Alcohol test consisted of four questions about the three types of alcoholic products: beer, wine and liquor.

For each alcoholic product, subjects were asked to indicate how often they usually drink, how much they usually drink, the most they have ever drank at one time and how often they have ever drank this maximum amount. After completing the two analyses, the result were correlated with each other and with alcohol consumption measures. They concluded from the two analysis procedures that there was little difference between latent trait analyses and the classical test analysis.

In considering the procedure and methodology of the study the authors failed to establish the unidimensionally, that is, the factors structure of the instruments. They also disregarded the issue of validity of the instruments by not exploring the fit of the latent trait model to the data. The study was conducted in the content of a realistic clinical concern, the primary focus of the present study was concerned with cognitive outcome in a physics achievement test, using the one parameter and the two- parameter latent trait models.

Vandarlinden (1981) investigated the pretest-posttest validation of a criterion-referenced test and the test information function of the latent trait model. An instructional unit of "forces and motions" from a course in physics in which 10th grade students, an equivalent of senior secondary one students were instructed in elementary mechanical concepts of mass, inertia, speed acceleration and laws of motion. Altogether 182 subjects participated in the investigation. The pretest-posttest design was realised by administering a 25 item four weeks before the first and shortly after the last class.

The pretest-post-test statistic designated Dpp was computed by subtracting the proportion of subjects who correctly answered the item on a pretest from the proportion of subjects who correctly answered the items on a post test. Data analysis were conducted using the one-parameter model on the BICAL computer programme. The overall fit of the items to the model was 1.026 (0.00 probability level) and was used to illustrate the differences between the pretest-posttest coefficient and the test information function. The correlations between the two statistics was 0.23 for one cut –off point and –0.19 for another cut-off point.

He concluded that the two statistics would lead to the selection of different subset of items. Although the study appears logical and empirical, the problem with the study however borders on the extent to which the physics test was unidimensional. The other problem with the study has to do with the sample size. The 182 subjects used in the study was too small to reliably estimate the test information function of the physics achievement test. The other problem with the study concerns mortality of sample after pretesting. The author failed to discern how this source of invalidity was control in his experimental study. While the study was for a criterion. referenced test, the present study was aimed at developing and establishing the psychometric properties of a physics achievement test using the one- parameter and the two- parameter latent trait models.

A comparison of three types of item analysis procedures using the classical and the latent trait models was investigated by Benson, Croker and Ware (1979). The instrument for the study was the verbal aptitude subtest. Classical item analysis, factor analysis and the Rasch model analysis was used in the development of 30 items of a verbal aptitude subset which was administered to high school students. The results of the analysis showed that there was no apparent difference in the three types of tests produced by the three methods of item analyses, with regard to the precision of measurement. The authors therefore concluded that there was no substantive differences in terms of test efficiency in the three item analysis procedures for students of varying ability levels.

One major problem with the study is that we cannot be sure that the conclusion applied to a formula- scored, verbal aptitude subtest, as the model deals with binary scores. Another problem with the study is that the authors failed to establish the single trait for ability of the verbal aptitude subject in conformity with the Rasch's model. While the study way based on comparing the precision and efficiency of three methods of item analyses using aptitude subtest, it was the purpose of the present study to develop and establish the psychometric properties of a physics achievement test using the one- parameter and the the two-parameter latent trait models. In a related study, Harwett (1983) performed an empirical study to establish the relationship between the pretest-posttest statistic, DPP and the latent trait statistic. When the two selection methods were compared across multiple data set, the latent trait statistics was generally found to produce more reliable test than those composed of items selected by the pretest-post-test statistic. However, the major problem with the study was the single trait for the data sets which the investigator failed to explore.

In order to extent the use of latent trait models across the full spectrum of mental tests, Reckase, (1979) conducted a study on the applicability of a uni-factor latent trait models to multi-factor tests using the one-parameter and the three-parameter latent trait models. In order to evaluate the effect of multidimensionality on the one-parameter and three-parameter models, Reckase obtained 10 data sets varying in the number of factors present. All the data sets were subjected to three factors analyses in order to determine the factor structure of the empirical data sets. After the factor analyses, each of the data sets was analysed using the one-parameter and the three parameter latent trait models. In addition, classical item analysis, ability estimates and the overall fit of the models were investigated;

The findings of the study show that the one-parameter and the three-parameter models estimate different abilities. He concluded that the data sets measures independent factors, although they estimate the first component, when it was large relative to the other factors. The major problem with this study is that the author did not highlight how the data set were scored in view of the fact that the one-parameter and the three parameter models deal with binary scored items. The study used the one-parameter and the three –parameter models on ten data sets. The purpose of the present study was to develop and validate a physics achievement test, using the one-parameter and the two-parameter models.

Goldman and Raju (1986) carried out a study using the one- and the two-parameter models to validate an Attitude Survey Scale (ASS) on a sample of 3,000 employees of some major American companies, from 1979-1981. The Attitude survey was designed to measure their attitude towards their work environment. The attitude survey consists of 78 items questionnaire. Each item or statement has three response alternatives. "Agree"? and "Disagree". To confirm to the requirements of latent trait models, favourable response were scored one and unfavourable and .?" were scored 0.

The results of the study show that for the one-parameter model and the two-parameter model, the person ability (θ), and the item difficulty (bi) were accurately estimated and the attitude scale fits the model. They concluded that the sample ability parameters estimated from the one-parameter model and the two- parameter model were highly correlated. Although the study appears empirical and logical, the problem with the study has to do with what constitutes ability and "difficulty" for attitude response which the authors failed to highlight. This study was in the affective domain objectives. What are the probable findings in the cognitive domain objectives using a physics achievement test?.

Tinsley and Dawis (1975) investigated the applicability of the one-parameter latent trait model to analogy test items. It was hypothesised that latent trait item easiness ratio is invariant with respect to the ability level of the validating sample and that the ability estimate is invariant with respect to the ability level of the validating sample. Data were obtained from four samples: college, high school students, civil service clerical employees and clients at the Minneapolis vocational rehabilitation centre. The analogy test contains items on word analogy. Picture analogy and symbol analogy. Two data sets were taken independently and submitted to item analysis.

The product- moment correlation co-efficient was calculated between the two sets using the one-parameter model. The item easiness estimates from the different validating samples were positively correlated. It was concluded that the latent trait item easiness ratios were stable or invariance with respect to the ability of the validated sample, and that the stability of the model item easiness ratios was related to the goodness of – fit to the model. A possible draw back of the study is that the authors neglected to determine the extent to which the analogy test items were unidimensional in confirming with the assumptions of the one-parameter latent trait model. Also, the scoring procedures employed in the study were not explicit as the model deals with binary scores.

Anderson, Kearny and Everette (1968) investigated the fit of the one-parameter model to a 45-item intelligence test for screening applicants to the Australian Army and Royal Navy. samples of 608 recruit applicants to the citizen military Force (CMF) and 874 for the Royal Navy (RAN) were used for the investigation. For the CMF sample, 30 items fit the one-parameter model at the 0.01 level of significance, and 25 items fit the model at the 0.05 level of significance. For the RAN sample, the corresponding findings were 22 items and 16 items respectively.

When the product-moment correlation co-efficient was computed between the item easiness estimates obtained from CMF and RAN, a correlation co-efficient of 0.958 was found. They concluded that the model item estimates are more stable when only items which fit the model are considered. In consideration of the design of the study, the authors failed to establish to what extent the intelligence test items were unidimensionall in conformity with the one-parameter model. The authors also failed to discern how the intelligence test items were scored in view of the fact that the one-parameter model deals with binary scores. The present study developed and established the psychometric properties of a physics achievement test, using the one –parameter and the two- parameter Latent trait models.

Fischer (1973) investigated the applicability of the linear logistic model to a set of elementary tasks from the domain of differential calculus of explicit functions as taught in the eleventh grade of secondary schools, an equivalent of senior secondary two in Austria. Altogether 36 functions were presented to a sample of 287 students. Only the correctness of the differentiation was judged. The statistical analysis was essentially an estimation of the difficulty parameter (bi) for each task. He concluded that the data fit the model sufficiently.

The problem with the study is that its generalisability is questionable because the sample size (287) was very small for the partitioning necessary in the model's control. Also, the 36 functions used in the study was too small for the partitioning necessary to reliably estimate the item parameters using the unconditional maximum likelihood approach. The author also failed to ascertain whether the set of problems was homogeneous in the sense of the one-parameter latent trait model, throughout the learning process. The prime purpose of the present study was to develop and standardize a physics achievement test for senior secondary three physics students using the one-parameter and the two-parameter Latent trait models.

Gallini (1983) applied the one-parameter latent trait model to investigate the fit of the observed performance on the Raven progressive matrics items to what was expected by the model. The standardized Raven progressive matrices were administered to 151 seventh grade students, an equivalent of Junior secondary school one students in Nigeria, from an urban middle school in the South East Carolina. The latent trait item analysis was conducted using the unconditional maximum likelihood procedure, UNCON, of the computer programme BICAL. Results show that the observed proportions and those expected to correctly answered the items for the different ability groups were significant and fit the model. Gallini concluded that more than one trait underlies test performance on the Raven.

Although the design of the study appears eloquent, the findings however raised some important issues about the appropriate use of the model. One important issue relates to the sample size. The analysis was based on 147 subjects and was too small for generalising the findings of the study. One would certainty need a larger sample to reliably establish the item statistics or the fit to the model. The second issue relates to the unidimensionality of the Raven items. A truly factor analysis would have been more sufficient, as against the point biserial correlation co-efficient used by the author to investigate the factor structure of the Raven items..

Bock and Lieberman (1970) applied the data for the Law School Admission Test (LSAT) for the estimation of items and person parameters using the unconditional maximum likelihood approach. The data represent responses of 1,000 subjects drawn from a sample of subjects who applied for admission into law schools at various universities in the United States. The data cited were sampled by means of a stratified sampling procedures with respect to university and level of achievement within universities.

The Law School Admission Test was a 60-item test. For example: items 11 to 30 of sections six and seven consist of highly homogeneous figure classification items and items on debate. The reliability co-efficients (KR-20) were 0.88 and 0.765

respectively. They concluded that there were good fittings between the data and the latent trait models. Although the study appears substantial in terms of design, the problem with the study has to do with the formula applied in scoring the items in view of the fact that the model deals with binary scores. This the authors failed to highlight. The other problem with the study relates to the single trait for ability of the law school admission test, which the authors failed to investigate. The present study developed and established the psychometric properties of a physics achievement test using the one-parameter and the two-parameter latent trait models.

Lord (1968) applied the conditional approach to the analysis of the verbal subtest of the scholastic Aptitude Test (SAT) using the three-parameter latent trait model. The study was a statistical analysis of the data of a sample of 2,882 examinees who took the test. The sample was drawn from a larger group of 5,000 examinees by selecting those with the highest and lowest maths scores.

The conclusion was that there was high correlation between the two ability measures and that the model fit the data. The problem with the study is that we cannot be sure that the conclusion applied to a formular scored items. Another drawback of the study relates to the unidimensionality of the scholastic Aptitude test, which the author failed to investigate. The present study developed and validated a physics achievement test using The one parameter and the two-parameter latent trait models.

John, Annabel, Ryan, Edens and Dixton (1995) investigated the factors derived from schools climate profile using the Rasch, one-parameter latent trait model. The school climate profile consists of eight subsets, each with a total of 40 items. For each item, respondents used one of the four descriptors: almost never = one; occasionally + two; frequently = three and almost = four. Three samples were used. The first sample consisted of 405 senior high school students. The second sample comprised 906 teachers, administrators, while the third sample consisted of 490 Junior school students.

Primary and Second order principal component analyses were used to examine the construct validity, the unidimensionality of the school climate factors. The Rasch analysis was used to test the fit for each sample, and the stability of the item parameters across samples. Based on the analyses they concluded that the school climate profile subtest lead to several inferences about the measurement of the school climate profile.

From the point of research, the study appears substantive. The methodological issues focused on whether or to what extent the factor analytic approach identified subsets of items that scale together using the Rasch is approach. This the authors failed to explore. The present study developed and validated cognitive out comes in physics for senior secondary school III physics students using the one-parameter and the two-parameter latent trait models.

Korashy (1995) applied the Rasch, one parameter measurement model to the selection of items for a mental ability test. The Otis-Lennon Mental Test, Advanced form J. was used for the study. The test consists of 80 items: 46 of which were verbal, 17 symbolic and 17 figural. Two hundred and sixty five students constituted the sample of the study. The MSCALE programme which utilises the PROX and UNCON procedures was employed to estimate the item parameters and person ability estimates.

Results indicated that using the classical test model procedures may be useful in the first stage of the test development, for revising the items. He concluded that the Rasch measurement model proved to be more effective in providing reliable and accurate parameters for item and person ability estimates. A possible limitation of the study is that the author failed to investigate the extent to which the intelligence test items were unidimensional in conformity with the assumptions of the Rasch model. Another drawback of this study was that the author failed to highlight the scoring procedures employed in the study, in view of the fact that the one parameter model deals with binary scored items.

Wright (1967) investigated the invariant of the Rasch item easiness ratios with respect to the ability level of the validation sample. The investigation was based on the answers of 976 beginning law students to 48 reading, comprehension items on the Law School Admission Test (LSAT). To obtain sample of different ability, Wright (1967) selected two contrasting groups from the total sample. The "dump" group included the 325 students who did poorest on the test while the "smart" group included the 303 students with the highest scores.

Wright compared the similarity between the two sets of the Rasch ability estimates and the two sets of the Percentile ranks. Based on the findings, Wright concluded that the Rasch model leads to sample-free test validation, while the classical test model does not. In considering the procedure and methodology of this study, the another failed to establish the unidimensionality, that is, the single trait for ability of the law school admission test. The author also disregarded the issue of validity of the law school admission test by not exploring the fit of the latent trait model to the data. The other problem with the study concerns the scoring procedures employed, as the model deals with binary scored items. The study employed the one parameter model on the data set. The purpose of the present study was to develop and establish the psychometric properties of a physics achievement test for senior secondary school III physics students, using the one-parameter and the two-parameter latent trait measurement models.

3. Summary of Literature Review

The literature review covered theoretical background of achievement testing and measurement in historical perspectives. The review literature explored the historical and conceptual background of the classical and latent trait models. Item analysis utilising both the classical and latent trait models were reviewed with regard to their capacity for provisional items selection. Within the context of the classical test model, it was argued that only factor analysis provides a rational procedure for item selection. In the context of latent trait models, two possible approaches for estimating item parameters were distinguished, conditional and unconditional maximum likelihood. The capacity for estimating item parameters solely from the patterns of item–by-subject data matrix is termed unconditional estimate. When the ability parameter is regarded fixed, it is termed conditional estimate.

Both theoretical and empirical evidence of instrumentation research were reviewed and the need for valid assessment instruments highlighted. Many authors based their studies on the classical test model. This is primarily due to its simplicity, both conceptually and computationally. Unfortunately, the classical test model has some serious disadvantages which made its use in developing and validating achievement tests questionable. For instance, in the classical test model, item statistics change from a sample whose mean ability is high to one who mean ability is low. In addition, the classical test model typically provides only one overall standard error of measurement for all items on the entire test. Because of the limitations of the classical test model, the present study was aimed at developing and establishing the psychometric properties of a physics achievement test, using the one- and two-parameter latent trait models. One problem with previous studies is that the authors failed to establish the unidimensionality, single trait for ability of the data they cited. Some investigators also disregarded the issue of validity of their instruments by not exploring the fit of the latent trait models to the data. Other problem with previous studies relate to small sample size, which limits the partitioning necessary in the model control. A further problem with previous studies concerns scoring procedures which were not explicit in view of the fact that latent trait models deal with binary scored items.

A further look at previous studies indicate that no empirical study has been investigated using a physics achievement test. In summary, existing problems in previous studies, as highlighted above, necessitated the present study which aimed at providing answers to the above problems utilising a physics achievement test developed and validated on a sample of senior secondary school III physics students, applying one-parameter and two-parameter latent trait models and the classical test model.

CHAPTER THREE

RESEARCH METHOD

This chapter concerns itself basically with the general plan for carrying out the study and procedures for data analysis, aimed at answering the research questions and testing the hypotheses formulated. Specifically, this chapter described the design of the study, the area of the study, population, sample and sampling techniques. In addition, instrumentation as well as methods for establishing its validity and reliability, procedures for data collection and item analysis are presented. Finally, latent trait models and computer programmes used in the study are discussed.

Research Design

This study has been designed to be an instrumentation research since it involves the development and validation of an instrument, a physics Achievement Test, which can be used in evaluating cognitive learning outcomes of Senior Secondary School Physics students. Ali (1996) defined instrumentation research as a study, which is geared towards the development and validation of measurement instruments in education. The International Centre for Education (ICEE, 1982) defines instrumentation research as a study, which aims at introducing new contents, procedures, technologies or instruments for educational practices.

Area of the Study

The area of the study was Rivers State, Nigeria. There are 23 Local Government Council areas in Rivers State, namely: AKULGA, ANOLGA,

ASALGA, DELGA, OGBOLGA, ONOLGA, WALGA, OLGA, ABOLGA, ALGA-EAST, ALGA-WEST, GOLGA, KELGA, OBALGA, ONELGA, PHALGA, TALGA, OYILGA, OMULGA. EMUOLGA, ELGA, ELELGA AND KHALGA. The first eight council areas are strictly Riverine, while the remaining 15 council areas are upland. There were 192 Senior Secondary Schools, Government Secondary in Rivers State during the 1998/99 academic year. These comprised 17 single boys, 20 single girls and 155 co-educational Senior Secondary Schools.

Population

The population of this study is made up of all Senior Secondary School III Physics students numbering 16, 109 who offered physics in Rivers State, during the 1998/99 academic year. Details of population are presented in Table 1.

Table 1

Local Government Areas and Population of the Study

S/NO	Local Govt. Area	Single Boys	Single Girls	Mixed	Total	Total Number Of Students
1.	PHALGA	4	3	4	11	2587
2.	OBALGA	2	3	6	11	1781
3.	EMUOLGA	-	1	14	15	725
4.	KELGA	1	1	10	12	1716
5.	ALGA-EAST	1	1	6	8	956
6.	ALGA-WEST	-	1	9	10	318
7.	ELGA	1	1	13	15	720
8	OMULGA	-	0	3	3	95
9.	ABUALGA	1	1	7	9	398
10.	ONELGA	1	1	9	11	643
11.	AKULGA	-	1	1	2	66
12.	OLGA	-	1	1	2	65
13.	DELGA	1	1	10	12	575
14.	ASALGA	1	1	6	8	293
15.	KHALGA	-	1	17	18	1952
16.	ONOLGA	-	-	2	2	52
17.	GOLGA	1	1	8	10	774
18.	ANOLGA	-	-	9	9	135
19.	ELELGA	1	-	2	3	543
20.	TALGA	-	-	9	9	135
21.	OYILGA	-	-	5	5	573
22.	WALGA	1	1	3	5	388
23	OBOLGA	-	-	2	2	86
	TOTAL	**17**	**20**	**155**	**192**	**16,109**

SOURCE: West African Examination Council (WAEC) Statistic: Senior School Certificate Examination (SSCE), May/June, 1999, Rivers State.

Sample and Sampling techniques

The sample of this study consists of 2215 (1349 males and 866 females) Senior Secondary School Physics students who sat for the Senior School Certificate Examination during the May/June 1999 school year in Rivers State. Subjects came from diverse socio-economic backgrounds indicated by the educational and occupational status of their fathers. Their ages ranged from 16-20 years with a mean of 18 years and standard deviation of 2.6 years. The sample came from a mix of rural and urban schools. Approximately 66.0 per cent (1452) of the sample came from urban schools in the state.

The sample of the study was composed through a multi-stage sampling techniques involving simple random sampling, stratified proportionate sampling and cluster sampling.

The simple random sample was achieved by using a table of random numbers. To accomplish this, all the local government areas in the state were numbered serially from 1 to 23. The sample was drawn, using the first two and three digits. While drawing the sample, a number which repeats itself or a number larger than the population of local government areas (23) was ignored and the process was continued until the required sample size (13) was composed.

The second stage involved a stratified proportionate sampling of Senior Secondary School from the 13 sampled Local Government areas of the study Firstly, all schools in the sampled local government areas were divided into three strata representing the three types of schools in the state. The three types of schools are: boys only (13); girls only (13), and co-educational schools (77) given a total of 103 schools. The schools were then drawn proportionate to the type from each stratum based on the ratio of the total number of schools in each school type. Applying the procedure, 12.62 percent or two schools, were drawn from boys only 12.62 percent or two schools were drawn from girls only, while 74.76 percent or 58 schools were drawn from co-educational schools, given a total of 62 Senior Secondary Schools of the study.

The third stage involved cluster sampling. All senior secondary III physics students in the 62 sampled schools were the subjects of this study. Since Senior Secondary School III physics class is an optional class normally selected by the more Science inclined students, the students constitute a self-selected group (Mckezie & Padilla, 1986). Employing this sampling procedure, each sampled school constitutes a cluster, a unit of study, hence all physics students in all the sampled schools were studied. This multi-stage sampling procedures were employed because they provide wider coverage, and is efficient in terms of the number of subjects drawn for the study. Details of sample drawn from each stratum are presented in Table 2.

Table 2

Sample Schools And Subjects of the Study

S/NO	NAME OF SCHOOL	LOCAL GOVT. AREA	CENTRE NUMBER	SCHOOL TYPE	MALE	FEMALE	TOTA
1.	Stella Maris College, Port Harcourt	PHALGA	3114	Single boys	63	-	63
2	Govt Girls Sec. School, Harbour Road, Port Harcourt.	PHALGA	31263	Single girls	-	82	82
3	Govt Sec. Sch. Borokiri, PH.	PHALGA	31259	Mixed	65	39	104
4	Community Sec. Sch. Nkpolu	PHALGA	31373	Mixed	15	10	25
5	Community Sec. Sch., Abuloma	PHALGA	31393	Mixed	11	19	30
6	Sports Institute, Isaka, PH	PHALGA	31281	Mixed	9	6	15
7	Govt Girls Sec. Sch. Rumuokuta	OBALGA	31224	Single girls	-	45	45
8	Community Sec. School, Okoro-Odo	OBALGA	31195	Mixed	44	28	72
9	Obio Comprehensive College	OBALGA	31236	Mixed	45	31	76
10	Govt Sec. Sch. Elekahia	OBALGA	31249	Mixed	21	11	32
11	Govt. Sec. Sch., Eneka	OBALGA	31314	Mixed	8	14	22
12	Comm. Sec. Sch., Rumukwurushi	OBALGA	31369	Mixed	53	21	74
13	Comm. Sec. School, Rumuapara	OBALGA	31382	Mixed	13	9	22
14	Govt. Sec. Sch. Ihugbogo	ALGA-EAST	31105	Mixed	35	13	48
15	Upata II Govt. Sec. Sch.,Edeoha	ALGA-EAST Q31	31106	Mixed	15	13	28
16	Govt. Sec. Sch., Ogbo	ALGA-EAST	31370	Mixed	23	7	30
17.	Comm. Sec. Sch. Okporoma	ALGA-EAST	31401	Mixed	19	13	32

Table 2

Sample Schools And Subjects of the Study

S/NO	NAME OF SCHOOL	LOCAL GOVT. AREA	CENTRE NUMBER	SCHOOL TYPE	MALE	FEMALE	TOTA
1.	Stella Maris College, Port Harcourt	PHALGA	3114	Single boys	63	-	63
2	Govt Girls Sec. School, Harbour Road, Port Harcourt.	PHALGA	31263	Single girls	-	82	82
3	Govt Sec. Sch. Borokiri, PH.	PHALGA	31259	Mixed	65	39	104
4	Community Sec. Sch. Nkpolu	PHALGA	31373	Mixed	15	10	25
5	Community Sec. Sch., Abuloma	PHALGA	31393	Mixed	11	19	30
6	Sports Institute, Isaka, PH	PHALGA	31281	Mixed	9	6	15
7	Govt Girls Sec. Sch. Rumuokuta	OBALGA	31224	Single girls	-	45	45
8	Community Sec. School, Okoro-Odo	OBALGA	31195	Mixed	44	28	72
9	Obio Comprehensive College	OBALGA	31236	Mixed	45	31	76
10	Govt Sec. Sch. Elekahia	OBALGA	31249	Mixed	21	11	32
11	Govt. Sec. Sch., Eneka	OBALGA	31314	Mixed	8	14	22
12	Comm. Sec. Sch., Rumukwurushi	OBALGA	31369	Mixed	53	21	74
13	Comm. Sec. School, Rumuapara	OBALGA	31382	Mixed	13	9	22
14	Govt. Sec. Sch. Ihugbogo	ALGA-EAST	31105	Mixed	35	13	48
15	Upata II Govt. Sec. Sch.,Edeoha	ALGA-EAST Q31	31106	Mixed	15	13	28
16	Govt. Sec. Sch., Ogbo	ALGA-EAST	31370	Mixed	23	7	30
17.	Comm. Sec. Sch. Okporoma	ALGA-EAST	31401	Mixed	19	13	32

S/NO	NAME OF SCHOOL	LOCAL GOVT. AREA	CENTRE NUMBER	SCHOOL TYPE	MALE	FEMALE	TOTA
18.	Comm. Sec. Sch.,Amaji	OMULGA	31145	Mixed	8	7	15
19.	Comm. Sec. Sch. Umuogba	OMULGA	31192	Mixed	7	3	10
20.	Govt. Sec. Sch., Eberi-Omuma	OMULGA	31312	Mixed	9	9	18
21.	Comm. Sec. Sch.,Anyu	ABUALGA	31177	Mixed	7	5	12
22.	Govt. Sec.Sch. Okpedan	ABUALGA	31247	Mixed	7	3	10
23.	Govt. Sec. Scn., Abua	ABUALGA	31280	Mixed	13	10	23
24.	Govt. Sec. Sch. Emelego	ABUALGA	31335	Mixed	9	-	9
25.	Community Sec. Sch., Egbolom	ABUALGA	31335	Mixed	8	7	15
26.	Comm. Sec. Sch. Otapha	ABUALGA	31338	Mixed	16	4	20
27.	Nyemoni Grammar Sch., Abonnema	AKULGA	31250	Mixed	12	8	20
28.	Bonny National Grammar Sch., Bonny	OLGA	31266	Mixed	11	9	20
29.	Comm. Sec. Sch., Bille	DELGA	31128	Mixed	26	15	41
30.	Comm. Sec. Sch. Tombia	DELGA	31222	Mixed	22	8	30
31.	Comm. Sec. Sch., Ke	DELGA	31380	Mixed	29	11	40
32.	Govt. Sec. Sch.,Usokun	DELGA	31380	Mixed	15	8	23
33.	Comm. Sec. Sch., Uegwere	KHALGA	31117	Mixed	12	9	21
34.	Comm. Sec. Sch., Kono-Boue	KHALGA	31122	Mixed	30	17	47
35.	Comm. Sec. Sch.,Kpean	KHALGA	31123	Mixed	60	27	87
36.	Comm. Sec. Sch, Wiiyaakara	KHALGA	31197	Mixed	15	12	27
37.	Birabi Memorial Grammar Sch. Bori.	KHALGA	31261	Mixed	14	17	31
38.	Govt. Sec.Sch., Kaani	KHALGA	31353	Mixed	50	31	81
39.	Govt. Sec. Sch., Kaa	KHALGA	31333	Mixed	32	11	43
40.	Kono Sec. Sch., Kono	KHALGA	31124	Mixed	15	9	24
41.	Beeri High Sch.,Beeri	KHALGA	31199	Mixed	8	2	10
42.	St. Pius X College, Bodo	GOLGA	31253	Single Boys	47	-	47
43.	Comm. Sec.Sch.,Bera	GOLGA	31120	Mixed	42	21	63
44.	Comm. Sec. Sch. K-Dere	GOLGA	31121	Mixed	15	12	27

S/NO	NAME OF SCHOOL	LOCAL GOVT. AREA	CENTRE NUMBER	SCHOOL TYPE	MALE	FEMALE	TOT
45.	Comm. Sec. Sch. Bua-Yeghe	GOLGA	31212	Mixed	19	14	33
46.	Comm. Sec. Sch.,Mogho	GOLGA	31227	Mixed	11	10	21
47.	Govt. sec. Sch., Kpor	GOLGA	31337	Mixed	23	13	36
48.	Comm. Sec.Sch., Bomu	GOLGA	31118	Mixed	13	7	20
49.	Comm. Sec. Sch., Biara	GOLGA	31119	Mixed	14	11	25
50.	Govt. Sec. Sch. Onne	ELELGA	31326	Mixed	27	12	39
51.	Comm. Sec.Sch., Alesa-Eleme	ELELGA	31221	Mixed	27	17	44
52.	Comm. Sec. Sch., Korokoro-Tai	TALGA	31152	Mixed	32	19	51
53.	Comm. Sec. Sch.,Sime Tai	TALGA	31205	Mixed	28	11	39
54.	Comm. Sec. Sec., Ban-Ogoi	TALGA	31106	Mixed	23	20	43
55.	Comm. Sec.Sch., Bunu-Tai	TALGA	31107	Mixed	22	19	41
56.	Govt. Sec.Sch., Kpite Tai	TALGA	31325	Mixed	11	10	21
57.	Comm. Sec. Sch., Koroma-Tai	TALGA	31403	Mixed	21	12	33
58.	Comm. Sec. Sch.,Kira-Tai	TALGA	31209	Mixed	15	11	26
59.	Comm. Sec. Sch., Botem-Tai	TALGA	31210	Mixed	14	9	23
60.	Comm. Sec. Sch., Okochiri	WALGA	31150	Mixed	26	12	38
61.	Comm. Sec. Sch., Kalio-Ama	WALGA	31151	Mixed	15	12	27
62.	Okrika National Sec. Sch., Okrika	WALGA	31204	Mixed	30	11	41
	TOTAL				**1349**	**866**	**2215**

Instrumentation

The instrument for this study consists of a 60-item physic achievement test selected from an initial pool of 87 items (Appendix 1), generated based on the Senior Secondary School Physics curriculum.

Development of the Physics Achievement Test (PAT)

The development of the physics Achievement Test (PAT) proceeded in a number of identifiable stages: planning, preparing, pilot testing and evaluating in line with suggestings by Anene & Ndubusi (1992); Herman (1990); Martuzza (1977); Wesman (1971). The planning state of the physics achievement test involved a review of Senior Secondary School Physics curricula and content analysis of physics text books such as Nelkon & Parker (1968); Chapple (1976); Okeke & Anyakoha (1987). Based on this analysis and interview, with physics teachers at the senior secondary school level, PAT outlines and objectives were formulated to reflect distributions in areas of knowledge, comprehension, application, analysis, synthesis and evaluation, following Bloom, Englehart, Furst, Hill and Krathwohl taxonomy of educational objectives for cognitive domains (1980), using the table of specification to ensure its content validity. The participation of the physics teachers increases the likelihood that the content, vocabulary as well as complexity of the PAT items were appropriate for the Senior Secondary School physics students. The teachers also attested that the students have been exposed to the contents of PAT prior to its administration. The second stages in the development of the PAT items involved the actual preparation, writing and ascertaining correct alternative options of the PAT items. Based on the test outlines, interview with physics teachers and content analysis of physics text book and curriculum, an initial pool of 87, multiple choice items with 5 alternative options PAT Items were generated (see appendix II) options PAT Secondary physics curriculum, namely: 1 matter, position, motion and time, II Energy (Mechanical and heat) III waves, IV fields, and V. Atomic/nuclear physics and electronics using the table of specification to ensure its validity (see table3)

Table 3

Table of specification for the physics Achievement Test (PAT)

Contents	Know-ledge 35%	Compre-hension 30%	Applica-tion 20%	Analysis 5%	Synthesis 5%	Evaluation 5%	Total 100%
I. 25%	5	5	3	1	1	1	16
II. 15%	3	3	2	0	0	0	8
III. 25%	5	5	3	1	1	1	16
IV. 15%	3	3	2	0	0	0	8
V. 20%	4	3	2	1	1	1	12
Total 100%	**20**	**19**	**12**	**3**	**3**	**3**	**60**

I = Matter, position, motion an time

II = Energy (Mechanical and heat

III = Waves

IV = Fields

V = Atomic/Nuclear physics and Electronics

The PAT items were checked with the table of specification to ensure that the test showed the desired proportion of emphasis. The weighting of the content areas was I: = 25 percent, II = 15 percent, III = 25 percent, IV = 15 percent and V = 20 percent. For the objective domain, the weighting were: knowledge = 35 percent, comprehension = 30 percent, application = 20 percent, analysis = 5 percent, syntheses = 5 percent and evaluation = 5 percent. The rationale for the differential weightings of the content areas and the objective domains was based on the relative emphasis on the content areas and the objective as outlined in the Senior Secondary School Physics curriculum, as well as on interviews the researcher had with a cross-section of physics teachers. The actual number of questions assigned to each cell was determined using the formula: weight of content in percent multiplied by weight of objective domain in percent multiplied by the total number of questions (Anene & Ndubusi, 1992).

All the PAT items of a particular content area were placed together in the test. This was an attempt to avoid a situation whereby varying number of items content were thrown together in a random order in the entire test. The instructions on test direction were made clear, complete and concise so that subjects would know what they were expected to do. A conscious effort was made to vary the position of correct options in order to prevent the occurrence of a systematic answer pattern. A chi-square goodness-of-fit statistic was used to test if there was a systematic bias in the choice of correct positions of options (see Appendix III). The obtained value, $x^2 = 3.33$ ($n = 5$, $df = 4$), was not significant, hence there was no statistically significant bias in the choice of correct of options, Options for items having numerical response alternative were arranged in ascending order of magnitude as suggested by Martuza (1977) and Henrysson (1971). Items were ordered according to their difficulty levels with easiest items appearing first and the most difficult last in order to motivate the subjects into the test and help minimise anxieties and frustrations which tend to accompany an encounter with difficult items at the beginning, the last item on any page ended on that page with no part appearing on the next. Background information including gender, age of subjects was obtained by a section of the test.

Initial validation of the physics Achievement Test (PAT)

The PAT items were submitted to three experts, practising physics teachers at the University of Nigeria, Nsukka and Rivers State Polytechnic, Bori for careful editing and critical review of the wording of the test items in order to avoid the inclusion of irrelevant, defective items and to establish the face validity of the PAT. Each expert was given a set of objectives and a copy of the test. To determine the PAT's content validity, each expert matched the PAT items to the appropriate domain. The experts also completed the PAT items so that the correct options could be ascertained. The number of correction answers selected by the experts in agreement with the answers selected by the researcher provided a measure of the objectivity of the test items. In addition, comments on objective, clarity and appropriateness of items were solicited to allow for modifications (see appendix iv). Following their comments, items were revised, modified or deleted. Specifically five items (nos. 11, 25, 27, 40 and 79) were deleted (see Appendix v).

Pilot Study

The purpose of the pilot study were to establish the test reliability (classical) and to compute item validity to aid in the test revision. The other purpose of the pilot study was to enable the investigator to gain insight into the usefulness, clarity and practicality of the proposed study. In addition, the pilot study was conducted in order to use the answers given as a basis for carrying out the distractor analysis. The first revised form of PAT (81 items) was administered to 240 (128 boys and 112 girls) Senior Secondary III Physics students following a standardised procedure in a pilot study. The results of the pilot study are presented in appendix ix.

Criteria for Item Selection

The following criteria were applied in the selection of items for inclusion in the final form of the physics achievement test.

1. The physics achievement test was limited to 60 items because this length makes it suitable for application within a classroom period of one hour in a school setting.
2. Only PAT items with a fit statistic not larger than 1.5 was included in the final form of PAT (Korashy, 1995; Bryce, 1981).
3. The standard error (SE) of the item validity are an indication of precision of measurement. The smaller the SE the greater its precision of measurement. Only PAT items with SE values that did not exceed 0.17, standard deviation (SD) of the item validity were included in the final form of PAT.
4. Distractor indices. The PAT items were selected if all wrong options possessed negative distractor indices.
5. The item difficulty values were taken into consideration to ensure a reasonable spread of items along the entire test. This is important because the measurement of the varied ability levels in the target population requires a similar variation in item difficulties.

Procedure for Data Collection

The resulting 60-PAT item was administered to the subjects by the investigator with the help of physics teachers in each of the sampled schools. In order to ensure objectivity of data collection and control for possible procedural bias, procedures for test administration were standardised. That is, to ensure uniform administration conditions across all schools, physics teachers involved in the data collection were given guidelines detailing specific instructions on how to administer the test. Subjects were assured that their scores would be classified by numbers and not according to their names. This was an attempt to protect the privacy of the subjects. The subjects were required to shade one of the symbols by each item that most appealed to them. The PAT items were scored correct or incorrect (1 or 0), and the final score was the total number of correct answers.

Method of Data Analysis

The following statistical analyses were employed to answer the research questions and to test the hypotheses that guided the study.

1. Test for uni-dimensionality. The extent to which the PAT items were uni-dimensional was investigated using factors analysis. Specifically, the default missing data option (listwise deletion) from the SPSS, for MS WINDOWS release 8.0 statistical package was used.
2. Parameter estimation. Item difficulty indices were estimated using the PROX Procedures as exemplified in the work of Izard & White (1980).
3. Parameters estimation. Item difficulty indices and discrimination indices of the two-parameter latent trait model were estimated applying regression technique the regression of an item on latent ability (Ross,1966).
4. Persons parameter estimation. The magnitude of the ability estimates and their associated standard errors wee estimated using the PROX procedures.

5. The standard error, reliability associated with item, was estimated using the PROX Procedures.

6. Analysis of fit. The chi-square goodness-of-fit test was used to investigate how well the latent trait model fit the PAT items (Anderson, 1973; Gallini, 1983). The chi-square test of fit involves a division of the subjects into ability subgroups by score levels. In the presents analysis, three groups (low, medium and high) were created on the basis of the raw scores.

7. The classical co-efficient were estimated using the Microsoft Excel computer programme.

8. Hypothesis one, the hypothesis of no significant relationship between the latent trait item difficulties, was investigated using the Pearson product moment correlation analysis. To obtain samples of different abilities, two contrasting groups designated "high" ability (Sample one) "low" ability (sample two) were selected from the total sample. The high ability group consisted of 488 subjects with higher scores, their lowest score being 40. The low ability group consisted of 923 subjects who did poorer the test, their highest score being 19. Evaluation of the strength of the correlation co-efficient was performed using the F-ratio9Leabo, 1972)

9. The Pearson product-moment correlation co-efficient was used to test hypothesis two.

10. Hypothesis three was investigated using techniques. To do this, item validities for each model: One-parameter model, two-parameter model and the classical test model were computed. These three measures were then inter-correlated to determine the relationship among each measure. Evaluated of the strength of these co-efficient was performed using the F -ratio (Leabo, 1972); two-way Analysis of Variance.

11. The Pearson product-moment correlation co-efficient was used to test hypothesis four.

Models and Computer Programmes

Two latent trait models were used in this study. They are the one-parameter and the two-parameter models. These models were selected because of their frequent use in the research literature and for their mathematical tractability. The validation procedures used in this study were selected after a review of six latent trait programmes written in BASIC, FORTRAN, BICAL, LOGISTS, FACON and MSCALE. The one-parameter latent trait programme selected was the BASIC. This analysis was performed using the PROX procedures on Microsoft Excel visual BASIC Computer programme. This programme computes several statistics for each item diffulty and summary statistics for all items and ability estimates. The two-parameter model programme selected model by setting C, the guessing parameter in the three-parameter model equal zero. These programmes were selected because of their accuracy and usability for the test length and sample size that was drawn for the study. These programmes estimated item difficulty, discrimination power and ability estimate and are written in FORTAP, FORTRAN Test Analysis Package (Baker & Martin, 1969).

The classical test model analysis was performed using the Excel computer programme. This programme computed the total score for each person and then correlated each item with the total score. These indices were designated the classical test version of the physics achievement test. The actual running of the programmes for data analysis was carried out at the Computer Centre of the Rivers State Polytechnic, Bori.

CHAPTER FOUR

Results

The purpose of the study was to develop and establish the psychometric properties of a physics achievement Test (PAT) on a sample of Senior Secondary III Physics student using the one-parameter and two-parameter latent trait models. In accordance with this purpose and the specific objectives of the study, seven research questions and four hypotheses were formulated to guide the study showed that :

1. Each item was attempted by a majority of the subjects and there was no item that all the subjects answered right or all the subjects answered wrong. This means that all the PAT items satisfied the requirement of latent trait measurement model (see Appendix ix).
2. The answer patterns to most of the items were reasonably distributed among the five alternative options. This indicates that the distracters were effective(see Appendix viii).
3. Three items (Nos. 7,13, 26) have positive distracter indices in one option. These alternative option would be modified.
4. The correct option attracted the highest percentage of scores on all items, indicating that the items were suitable for abilities in a meaningful fashion, they are reported separately for each research question or hypothesis formulated. This mode of presentation was selected so that the meaning of the results could be presented clearly.

Research Question One:

How Unidimensional, single trait for ability is the physics achievement test items?

Factor analysis was used to establish the unidimensionality, single trait for ability of the physics achievement test. The factors was on R-type where pairs of the PAT items were correlated using the principal component technique. A varimax rotation, with Kaiser Normalization, was performed on all factors, and iterated to a satisfactory solution, convergence.

The Eigen values, percent of variance and percent of cumulative variance, are presented in Table 4.

Table 4:

Results of Factor Analysis (Varimax Rotation with Kaiser normalization)

Component	Eigen values	Percent of variance	Percent of cumulative variance
1	14.811	25.104	25.104
2	2.315	3.924	29.028
3	1.720	0.915	31.943
4	1.536	2.603	34.546
5	1.520	2.577	37.123
6	1.454	2.465	39.587
7	1.312	2.224	41.811
8	1.287	2.181	43.992
9	1.174	1.990	45.982
10	1.142	1.935	47.917
11	1.114	1.888	49.805
12	1.070	1.814	51.619
13	1.026	1.739	53.358
14	1.015	1.720	55.079
15	0.991	1.679	56.079
16	0.925	1.568	58.325
17	0.922	1.563	59.888
18	0.877	1.486	61.374
19	0.847	1.435	62.809
20	0.837	1.418	64.809
21	0.823	1.395	65.622
22	0.804	1.363	66.985
23	0.777	1.317	68.302
24	0.771	1.308	69.609
25	0.756	1.282	70.891
26	0.740	1.253	1.72.145
27	0.725	1.229	73.373
28	0.704	1.194	74.567
29	0.691	1.170	75.738

Component	Eigen values	Percent of variance	Percent of cumulative variance
30	0.685	1.160	76.898
31	0.660	1.118	78.016
32	0.648	1.099	79.115
33	0.646	1.094	80.209
34	0.631	1.070	81.279
35	0.631	1.039	82.318
36	0.588	0.996	83.314
37	0.564	0.955	84.270
38	0.550	0.933	85.202
39	0.540	0.916	86.118
40	0.523	0.887	87.862
41	0.506	0.857	87.862
42	0.501	0.850	88.712
43	0.486	0.825	89.537
44	0.481	0.815	90.352
45	0.459	0.777	91.900
46	0.455	0.771	91.900
47	0.447	0.758	92.659
48	0.432	0.732	93.391
49	0.420	0.711	94.102
50	0.412	0.698	94.800
51	0.396	0.672	95.472
52	0.375	0.635	96.106
53	0.372	0.6031	96.737
54	0.355	0.602	97.339
55	0.348	0.590	97.929
56	0.325	0.551	98.480
57	0.313	0.531	99.011
58	0.308	0.522	99.532
59	0.276	0.468	100.000

The results presented in Table 4 show that a factor analysis of the physics achievement test resulted in 14 Eigen values greater than 1.0. The first Eigen value, 14.811 accounted for 25.184 percent of the variance. This satisfied Reckase's (1979) minimum criterion (24.00-36.00) for establishing the unidimensionality of the physics achievement test. This means that a single latent ability is necessary to adequately account for an examinee's performance on the physics achievement test and that the test measures a single trait. This further lends credence that the physics achievement test validity measures one broad factor called general achievement in physics.

Research Question Two

What are the estimates of the item parameter; (difficulty indices) using the one - parameter latent trait model?

The results of the analysis of data for research question two are presented in Table 5

Table 5

Latent trait Model item Parameter for the 60 PAT items Estimated using PROX Algorithm Techniques

ITEM	ITEM SCORE	DIFFI-CULTY	STD. ERROR	ITEM	ITEM SCORE	DIFFI-CULTY	STD. ERROR
1	1598	-1.49	0.0578	31	1004	-0.10	0.0520
2	882	0.17	0.0529	32	935	0.05	0.0524
3	905	0.12	0.0527	33	951	0.02	0.0523
4	1016	-0.13	0.0520	34	902	0.13	0.0527
5	846	0.26	0.0533	35	1070	-0.25	0.0547
6	1126	-0.37	0.0518	36	749	0.49	0.0547
7	1053	-0.21	0.0519	37	867	0.21	0.0531
8	1087	-0.29	0.0518	38	925	0.07	0.0525
9	975	-0.04	0.0522	39	1005	-0.10	0.0520
10	1055	-0.22	0.0518	40	914	0.10	0.0526
11	1050	-0.20	0.0519	41	909	0.11	0.0526
12	940	0.04	0.0524	42	880	0.18	0.0529
13	1000	-0.09	0.0520	43	815	0.33	0.0537
14	984	-0.06	0.0521	44	1009	-0.11	0.0520
15	1015	-0.13	0.0520	45	960	0.00	0.0523
16	868	0.20	0.0530	46	756	0.47	0.0546
17	840	0.27	0.0534	47	886	0.16	0.0529
18	924	0.08	0.0525	48	939	0.04	0.0524
19	898	0.14	0.0527	49	1050	-0.20	0.0519
20	883	0.17	0.0529	50	916	0.09	0.0526
21	807	0.35	0.0538	51	860	0.22	0.0531
22	840	0.27	0.0534	52	940	0.04	0.0524
23	1007	-0.11	0.0520	53	859	0.23	0.0531
24	1268	-0.69	0.0523	54	845	0.26	0.0533
25	977	-0.04	0.0522	55	831	0.29	0.0535

ITEM	ITEM SCORE	DIFFI-CULTY	STD. ERROR
26	1248	-0.64	0.0522
27	817	0.32	0.0537
28	1108	-0.33	0.0518
29	1047	-0.20	0.0519
30	1045	-0.19	0.0519

ITEM	ITEM SCORE	DIFFI-CULTY	STD. ERROR
56	857	0.23	0.0532
57	822	0.31	0.0536
58	932	0.06	0.0525
59	964	-0.01	0.0522
60	1066	-0.24	0.0518

Table 5, gives the latent trait model item difficult estimates for each of the 60 PAT items. Latent trait item difficult indices are expressed in log units with a mean of zero and standard deviation of one. A negative logit difficulty estimate corresponds to items having high difficulty indices (easy items), and positive logit difficulty estimate corresponds to items having a low difficulty indices (hard items). Result show that the items ranged in difficulty from -1.49, assigned to item one, an exceedingly easy item, to an index of 0.49, assigned to item 36, a some what difficult item. For the most part, the estimated values show consistency with the way the PAT items were written to increase in difficulty within each content area. Approximately 22 of the items or 36.67 per cent were easy with a difficulty level of less than zero. About 37 of the items or 61.67 were difficulty with a difficulty level of more than zero. One item or 1.67 per cent was of a medium difficulty level of more than zero.

The mean of the difficulty estimates was zero, standard deviation equal to 0.31, suggesting that there was little variability in score among the subjects. The difficulty estimates are, therefore, reasonably appropriate to the abilities of senior secondary physics students.

Research Question Three

What are the estimates of item parameters (difficulty and discriminating indices) using the two - parameter latent trait model?

This research question was answered using a regression technique involving the regression of an item on latent ability. In order to carry out this analysis, the proportion of examinees passing the item, p value was transformed into the corresponding z value read from a table of normal curve.

This transformation from p to z Provides a scale value which is linearly related to ability. Table 6 shows the estimated item parameters, difficulty and discrimination indices for the 60 PAT items.

Table 6

Item Parameter (Difficulty and Discrimination Indices Estimated from a sample (N =2215)Using the Two-Parameter Latent Trait model

Item	Proportion passing	Normal Deviate	Regression of item on latent ability	Discriminating indices	Difficulty indices
1	0.72	0.3372	0.523	1.04	-1.6
2	0.40	0.3867	0.720	1.76	0.95
3	0.41	0.3885	0.686	1.60	0.91
4	0.46	0.3970	0.515	1.02	0.77
5	0.38	0.3814	0.525	1.05	0.76
6	0.51	0.3988	0.413	0.77	-0.74
7	0.48	0.3984	0.376	0.69	0.73
8	0.49	0.3989	0.303	0.49	0.71
9	0.44	0.3945	0.439	0.83	0.75
10	0.48	0.3984	0.439	0.83	0.75
11	0.47	0.3980	0.222	0.39	0.69
12	0.42	0.3910	.0534	1.07	0.79
13	0.45	0.3961	0.400	0.74	0.73
14	0.44	0.3945	0.345	0.62	0.71
15	0.46	0.3970	0.380	0.70	0.73
16	0.39	0.3847	0.522	1.04	0.77
17	0.38	0.3814	0.541	1.09	0.77
18	0.42	0.3910	0.483	0.94	0.76
19	0.41	0.3894	0.484	0.94	0.76

Item	Proportion passing	Normal Deviate	Regression of item on latent ability	Discriminating indices	Difficulty indices
20	0.40	0.3867	0.510	1.01	0.76
21	0.36	0.3894	0.484	0.94	0.76
22	0.38	0.3814	0.519	1.03	0.76
23	0.45	0.3961	0.372	0.68	0.73
24	0.57	0.3925	0.292	0.54	-0.70
25	0.44	0.3945	0.522	1.04	0.79
26	0.56	0.3939	0.292	0.54	-0.70
27	0.37	0.3778	0.620	1.34	0.82
28	0.50	0.3989	0.472	0.91	-0.77
29	0.47	0.3980	0.491	0.96	0.78
30	0.47	0.3980	0.577	1.20	0.83
31	0.45	0.3961	0.502	0.99	0.77
32	0.42	0.3910	0.505	0.99	0.77
33	0.43	0.3932	0.420	0.79	0.74
34	0.41	0.3894	0.534	1.07	0.78
35	0.48	0.3984	0.388	0.72	0.74
36	0.34	0.3668	0.439	0.83	0.69
37	0.39	0.3847	0.446	0.85	0.73
38	0.42	0.3910	0.388	0.72	0.72
39	0.45	0.3961	0.506	1.00	0.78
40	0.41	0.3894	0.532	1.06	0.78
41	0.41	0.3894	0.526	1.05	0.78
42	0.40	0.3867	0.569	1.18	0.80

Item	Proportion passing	Normal Deviate	Regression of item on latent ability	Discriminating indices	Difficulty indices
43	0.37	0.3778	0.550	1.12	0.77
44	0.46	0.3970	0.527	1.05	0.79
45	0.43	0.3932	0.505	1.00	0.76
46	0.34	0.3668	0.613	1.32	0.79
47	0.40	0.3867	0.534	1.07	0.78
48	0.42	0.3910	0.498	0.96	0.77
49	0.47	0.3980	0.504	1.00	0.78
50	0.41	0.3894	0.580	1.21	0.81
51	0.39	0.3847	0.514	1.02	0.76
52	0.42	0.3910	0.538	1.09	0.79
53	0.39	0.3847	0.533	1.16	0.79
54	0.38	0.3814	0.533	1.07	0.77
55	0.38	0.3814	0.600	1.28	0.81
56	0.39	0.3847	0.456	0.87	0.74
57	0.37	0.3778	0.563	1.16	0.78
58	0.42	0.3910	0.570	1.18	0.81
59	0.44	0.3945	0.591	1.25	0.83
60	0.48	0.3984	0.569	1.18	0.82

Result show that the items are moderately difficulty and have uniform discrimination indices ranging from 1.76 to 0.39 (X = 0. 98, SD = 0.24). Item difficulty indices ranged from - 1.66, assigned item one, a very easy item to an index of 0.69, assigned items 11 and 36, a some what difficult items. An effective discriminating power consists of items with discriminating indices greater than 0.8 and an item difficulty index displaying a rectangular distribution from -2.0 to 2.0 for the latent trait two-parameter model (Urry, 1974). The estimated item parameter, therefore falls within the recommended rage. The average values of 0.98 for the discriminating power and 0.63 for the difficulty indices imply that the physics achievement test is valid and reliable.

Research Question Four

What are the estimates of the subjects measures (ability estimates) and their associated Standard error indicated by the raw scores on the physics achievement test?

Standardising a physics achievement test using latent trait models results in a logarithmic ability estimates being assigned to every possible raw score from 1 to n - 1 (n = number of items). The logits are interval units of measurement corresponding to total raw scores that have undergone an exponential transformation. The logits typically ranged from -4 to +4, positive logit means that the subject is relatively able on the test or unable (negative logit). Details are presented in Table 7.

TABLE 7

Log Ability Estimates and Associated Standard Errors (n= 2215)

RAW SCORE	COUNT	ABILITY ESTIMATE	STD. ERROR	RAW SCORE	COUNT	ABILITY ESTIMATE	STD. ERROR
1	4	-4.14	1.02	31	20	0.07	0.26
2	5	-3.42	0.73	32	44	0.14	0.26
3	5	-2.99	0.6	33	47	0.2	0.26
4	17	-2.68	0.53	34	53	0.27	0.26
5	16	-2.44	0.47	35	42	0.34	0.27
6	36	-2.23	0.44	36	38	0.41	0.27
7	36	-2.23	0.44	37	39	0.48	0.27
8	46	-1.9	0.39	38	33	0.56	0.27
9	74	-1.76	0.37	39	29	0.63	0.28
10	59	-1.64	0.35	40	16	0.7	0.28
11	103	-1.52	0.34	41	32	0.78	0.28
12	73	-1.41	0.33	42	32	0.86	0.29
13	80	-1.31	0.32	43	28	0.94	0.29
14	94	-1.21	0.31	44	42	1.03	0.3
15	50	-1.12	0.3	45	47	1.12	0.31
16	69	-1.03	0.3	46	27	1.21	0.31
17	37	-0.94	0.29	47	57	1.31	0.32
18	70	-0.86	0.29	48	38	1.41	0.33
19	42	-0.78	0.28	49	54	1.52	0.34
20	51	-0.7	0.28	50	53	1.64	0.35
21	55	-0.63	0.28	51	17	1.76	0.37
22	28	-0.56	0.27	52	7	1.9	0.39
23	47	-0.48	0.27	53	8	2.06	0.41
24	33	-0.41	0.27	54	2	2.23	0.44
25	40	-0.34	0.27	55	12	2.44	0.47
26	51	-0.27	0.26	56	6	2.68	0.53
27	27	-0.2	0.26	57	4	2.99	0.6
28	35	-0.14	0.26	58	4	3.42	0.73
29	44	-0.07	0.26	59	2	4.14	1.02
30	48	0	0.26				

The results presented in Table 7 shows that ability estimates ranged from -4.14, assigned to a raw score of one, an exceeding low score, to an estimate of 4.14, assigned to a raw score of 59, a substantially high score. Approximately 40.23 percent or 891 subjects were relatively able on the physics achievement test, logits ranged from 0 to 4.14. About 59. 77 percent or 1324 subjects were relative less able, logits ranged from -4. 14 to 0.27. The mean, of one ability estimates was -0.33, standard deviation equal to 1.18. This reflects a general low performance on the physics achievement test.

The standard error associated with ability estimates ranged from 0.26 to 1.02 (x = 0.3,SD = 0.07). The standard error indicates how much fluctuation would be like in a repeated estimated estimation of scores or ability estimates for the same examinee. The mean of the standard error associated with ability estimates is 26 percent of the standard deviation value of the ability estimate, implying that 74 percent of the total variance associated with ability estimates could be attributed to true variance.

Research Questions Five

What are the estimates of the standard error (reliability) for each of the physics achievements test items ?

Table 5 shows the standard error (reliability) for the physics achievement test items expressed in logit units. The standard errors ranged from 0.0578

to 0.0518 (x =0.0527, SD =0.0010). the mean of the standard error is o.17 or 17 percent of the total variance, unreliability, while 83 percent is attributed or due to true variance, reliability. This is a good precision of the physics achievement test. The largest standard error is less than 10 percent of the range of the standard error values. This means that the difficulty indices have been estimated with excellent precision.

RESEARCH QUESTION SIX

To what extent do the physics achievement test items fit the one parameter latent trait model?

The chi - square goodness -of - fit test was used to investigate how the latent trait model (one - parameter model) fit the PAT items. The chi - square test of fit involves a division of the subjects into ability subgroups by score levels. In the present analysis, three groups (low, medium, and high) were created on the basic of he raw scores. That is 27 per cent low, 46 per cent medium and 27 per cent high. The results presented in Table 8.

Table 8

Chi -square Goodness - 0f - Fit Test For the 60 PAT Items

	Observed Proportions of Subjects answering an item correctly			Expected Proportions of Subjects answering an item correctly			Chi-Sq.
ITEMS	**LOW 0-19**	**MEDIUM 20-39**	**HIGH 40-60**	**LOW 0-19**	**MEDIUM 20-39**	**HIGH 40-60**	
1	0.62	0.73	0.47	0.39	0.53	0.22	0.49
2	0.08	0.50	0.44	0.01	0.25	0.19	1.41
3	0.12	0.47	045	0.01	0.22	0.20	1.36
4	024	0.51	0.41	0.06	0.26	0.17	1.16
5	0.17	0.42	0.38	0.03	0.18	0.14	1.41
6	0.32	0.56	0.41	0.06	0.26	0.17	1.16
7	0.27	0.56	0.38	0.07	0.31	0.15	1.11
8	0.33	0.53	0.39	0.11	0.28	0.15	1.05
9	0.23	0.50	0.39	0.05	0.25	0.16	1.22
10	0.23	0.58	0.41	0.05	0.34	0.17	1.12
11	0.31	0.53	0.36	0.10	0.28	0.13	1.10
12	0.18	0.47	0.43	0.03	0.22	0.18	1.28
13	0.24	0.50	0.41	0.06	0.25	0.17	1.18
14	0.29	0.44	0.39	0.09	0.19	0.15	1.19
15	0.27	0.46	0.43	0.07	0.21	0.18	1.15
16	0.17	0.43	0.40	0.03	0.18	0.16	1.38
17	0.15	0.41	0.41	0.02	0.28	0.14	1.36
18	0.22	0.46	0.38	0.05	0.21	0.15	1.28
19	0.13	0.53	0.38	0.02	0.28	0.14	1.36
20	0.17	0.43	0.42	0.03	0.18	0.17	1.36
21	0.14	0.41	0.37	0.02	0.17	0.14	1.47
22	0.16	0.40	0.40	0.03	0.16	0.16	1.42
23	0.26	0.51	0.38	0.07	0.26	0.15	1.16
24	0.41	0.60	0.45	0.17	0.36	0.20	0.82
25	0.20	0.49	0.43	0.04	0.24	0.19	1.22
26	0.40	0.57	0.46	0.16	0.32	0.21	0.84
27	0.12	0.39	0.43	0.01	0.15	0.18	1.48
28	0.25	0.61	0.41	0.06	0.38	0.17	1.05
29	0.27	0.51	0.43	0.07	0.26	0.18	1.11
30	0.19	0.57	0.44	0.04	0.33	0.20	1.15
31	0.17	0.58	0.41	0.03	0.34	0.17	1.21
32	0.18	0.47	0.43	0.03	0.22	0.18	1.29
33	0.24	0.43	0.42	0.06	0.18	0.18	1.25

ITEMS	Observed Proportions of Subjects answering an item correctly			Expected Proportions of Subjects answering an item correctly			Chi-Sq,
	LOW 0-19	**MEDIUM 20-39**	**HIGH 40-60**	**LOW 0-19**	**MEDIUM 20-39**	**HIGH 40-60**	
34	0.11	0.53	0.41	0.01	0.28	0.17	1.37
35	0.28	0.54	0.41	0.08	0.29	0.17	1.08
36	0.13	0.39	0.34	0.02	0.15	0.12	*1.56
37	0.18	0.43	0.38	0.03	0.18	0.15	1.37
38	0.25	0.41	0.40	0.06	0.17	0.16	1.28
39	0.20	0.54	0.42	0.04	0.29	0.16	1.19
40	0.17	0.45	0.43	0.03	0.20	0.18	1.32
41	0.16	0.45	0.43	0.03	0.21	0.18	1.33
42	0.15	0.43	0.43	0.02	0.18	0.18	1.37
43	0.13	0.38	0.41	0.02	0.15	0.17	1.47
44	0.20	0.54	0.43	0.04	0.29	0.18	1.19
45	0.17	0.54	0.40	0.03	0.30	0.16	1.26
46	0.11	0.36	0.40	0.01	0.13	0.16	*1.57
47	0.17	0.43	0.42	0.03	0.18	0.18	1.36
48	0.17	0.51	0.40	0.03	0.26	0.16	1.29
49	0.20	0.56	0.44	0.04	0.31	0.19	1.13
50	0.16	0.46	0.43	0.03	0.21	0.19	1.32
51	0.13	0.48	0.39	0.02	0.23	0.15	1.41
52	0.16	0.50	0.42	0.03	0.25	0.18	1.29
53	0.15	0.42	0.42	0.02	0.17	0.18	1.40
54	0.15	0.42	0.41	0.02	0.17	0.17	1.42
55	0.08	0.47	0.41	0.01	0.22	0.16	1.47
56	0.17	0.43	0.38	0.03	0.19	0.14	1.39
57	0.12	0.40	0.42	0.01	0.16	0.18	1.47
58	0.16	0.46	0.45	0.03	0.24	0.19	1.25
59	0.18	0.49	0.44	0.03	0.24	0.19	1.25
60	0.19	0.60	0.44	0.04	0.36	0.19	1.13
GROUP COUNT	923	804	488				
MEAN	0.20	0.49	0.41				
STD. DEV.	0.09	0.07	0.02				
CHI SQUARE	75.48						

*Mis-fitting items

Table 8 shows the proportion of subjects answering an item correctly compared to the proportion expected to answer he item correctly. Results show that only two (or three percent) out of the 60 PAT items (nos. 36 and 46) had a fit statistic of more than1.50, meaning that these items did not fit the model. These items would be modified. The remaining 58items (0r 97 percent) had fit statistic ranging from 0.49 to .148 (x = 1.25, SD= 1.8). The results indicate that the chosen items had a good fit to the model and that the frequency distribution (observed scores) is close to the theoretical frequency distribution (expected scores). This means that the physics achievement test is compatible with the model. The mean difficulty, standard deviation and the group counts appear below the table.

Research question seven

What are the estimates of the classical test model item parameters, difficulty indices and point biserial correlation co-efficient

The results of the classical test model item analysis including the easiness estimates and the point biserial correlation of each item are presented in Table 9.

TABLE 9

Classical Test Model Item Analysis

	ITEM SCORE	DIFFI-CULTY	POINT BISERIAL		ITEM SCORE	DIFFI-CULTY	POINT BISERIAL
1	1598	0.72	0.23	31	1004	0.45	0.52
2	882	0.40	0.63	32	935	0.42	0.52
3	905	0.41	0.59	33	951	00.43	0.45
4	1016	0.46	0.45	34	902	0.41	0.55
5	846	0.38	0.46	35	1070	0.48	0.43
6	1126	0.51	0.41	36	749	0.34	0.43
7	1053	0.48	0.38	37	867	0.39	0.45
8	1087	0.49	0.34	38	925	0.42	0.40
9	975	0.44	0.44	39	1005	0.45	0.49
10	1055	0.48	0.43	40	914	0.41	0.51
11	1050	0.47	0.32	41	909	0.41	0.52
12	940	0.42	0.51	42	880	0.40	0.56
13	1000	0.45	0.43	43	815	0.37	0.52
14	984	0.44	0.38	44	1009	0.46	0.51
15	1015	0.46	0.43	45	960	0.43	0.50
16	868	0.9	0.49	46	756	0.34	0.54
17	840	0.38	0.50	47	886	0.40	0.51
18	924	0.42	0.42	48	939	0.42	0.51
19	898	0.41	0.49	49	1050	0.47	0.52
20	883	0.40	0.50	50	916	0.41	0.55
21	807	0.36	0.50	51	860	0.39	0.51
22	840	0.38	0.50	52	940	0.42	0.54
23	1007	0.45	0.40	53	859	0.39	0.52
24	1268	0.57	0.35	54	845	0.38	0.50
25	977	0.44	0.52	55	831	0.38	0.59
26	1248	0.56	0.37	56	857	0.39	0.46
27	817	0.37	0.58	57	822	037	0.56
28	1108	0.50	0.47	58	932	0.42	0.55
29	1047	0.47	0.48	59	964	0.44	0.54
30	1045	0.47	0.55	60	1066	0.48	0.54

Reliability (KR-20) = 0.89

The results show that:

1. The total raw score ranged from 1 to 59 (x = 25.97, SD = 14.19) The overall reliability (KR-20) was 0.89, implying that 89 percent of the total variance on the physics achievement test could be attributed to true variance. The true variance is the variance explained or association with differences in level of achievement.
2. The item easiness estimates ranged from 0.72 to o.32(x = 0.43, SD =0.06). About 8.33 percent of the items, 5 item, were easy (easiness level greater than 0.50), and about 91 .70 percent or 55 items were of medium difficulty easiness level ranged from 0.32 to 0.49. This means that the physics achievement test was suitable for the mental abilities of the target population of senior secondary school physics students which are assumed to be normally distributed.
3. The point biserial correlation (r pbs) between each item and the total test score, considered to be an indication of item validity in the classical test model, ranged from 0.23 to 0.63 (x = 0.48, 5D = 0.07) The above estimates are high and are within the recommended range (Youngman, 1979). This indicates that the physics achievement test items are highly homogeneous and internally consistent.

An additional analysis was conduced in which the physics achievement tet was broken down into different content areas/sub-test. Table 10 gives descriptive data, by raw score, for the various sub-test, while Table 11 gives the results of the Intercorrelation of scores on the different content areas of the physics achievement test.

Table 10

Means(X), Standard Deviation(SD) reliability co-efficient (KR20) for The various content Areas of the physics Achievement Test

Contents	No. of Items	MEAN	SD	KR20
1. Whole test	60	25.97	14.19	0.89
2. Time, positive, motion and matter	16	7.40	3.95	0.76
3. Energy (Mechanical and heat)	8	3.37	2.24	0.63
4. Waves	16	7.03	4.16	0.79
5. Fields	8	3.25	2.50	0.70
6. Atomic/Nuclear Physics and Electronic	12	4.94	3.63	0.79

The reliability for the items on time, position, motion and Matter was 0.76 (x =7.40, SD=3.950. The eight items on energy displays a reliability co-efficient of 0.63 (x=3.37 SD=2.24). The reliability for the 16 items on waves was 0.79 9 (x =7.03,SD=4.19) While the eight items on fields exhibited reliability co-efficient of 0.70 (x =3.25,SD = 2.50). The 12 items on Atomic/nuclear physics and Electronics displayed a reliability co-efficient of 0.70 (x =4.94, SD=3.63). The result indicates that the physics achievement test was reliable and had a good structural validity across the entire test. Table 11 contains the inter-correlations among the various content areas of the physics achievement test.

Table 11

Matrix of Intercorrelations of Scores on the Various Content Areas of the PAT (n = 60)

	Variable 1	*Variable 2*	*Variable 3*	*Variable 4*	*Variable 5*
Variable 1	1				
Variable 2	0.63	1			
Variable 3	0.73	0.68	1		
Variable 4	0.62	0.58	0.69	1	
Variable 5	0.64	0.64	0.71	0.68	1
TOTAL	16400	7467	15564	7154	10942
MEAN	7.40	3.37	7.03	3.23	4.94
STD.DEV	3.95	2.24	4.19	2.50	3.63

Variable 1 = Time, Position, Motion and Mater
Variable 2 = Energy (Mechanical and Heat)
Variable 3 = Waves
Variable 4 = Fields
Variable 5 = Atomic/Nuclear Physics and Electronics

The intercorrelations are all positive and range from 0.58 to 0.73, being highest for variables 1 and 3, and lowest for the pair 2 and 4. This means that the physics achievement test contains homogeneous items of similar content for each of the sub-test, as well as for the total test. The fact that all correlations are positive indicate that subjects who scored above or below average on any sub-test tended also to score above or below average on the other sub-tests. This is a reflection of the overall average performance on the physics achievement test, In addition, the point biserial correlation co-efficient for each item and sub-test score ranged from 0.42 to 0.62. This means that the PAT items are internally consistent and interrelated.

Hypothesis one

> There will be no significant relationship between latent trait model item difficulties obtained from two contrasting ability groups of the same Standardisation sample ($p > 0.05$)

This hypothesis was tested using the Pearson Product Moment Correlation coefficient. To obtain samples of different abilities, two contrasting groups designated "high" ability (Sample one) "and low" ability (Sample two) were selected from the total sample. The high ability group consisted of 488 subjects with higher scores, their lowest score being 40. The low ability group consisted of 923 subjects who did poorer in the test. Their highest score being 19.

Table 12

Difficulty Indices for 60 PAT items Computed independently in Two Samples:

SAMPLE 1 (N=488)			
ITEM	ITEM SCORE	EASI-NESS	STD ERROR
1	435	-2.57	0.08
2	403	-1.90	0.08
3	413	-2.08	0.08
4	382	-1.56	0.08
5	351	-1.15	0.08
6	383	-1.58	0.08
7	355	-1.20	0.08
8	359	-1.25	0.08
9	364	-1.31	0.08
10	375	- 1.46	0.08
11	333	-0.93	0.08
12	393	-1.73	0.08
13	377	-1.49	0.08
14	360	-1.26	0.08
15	395	-1.76	0.08
16	368	-1.37	0.08
17	374	-1.45	0.08
18	353	-1.17	0.08
19	349	-1.12	0.08
20	386	-1.62	0.08
21	342	-1.04	0.08
22	368	-1.37	0.08
23	355	-1.20	0.08
24	411	-2.04	0.08
25	401	-1.86	0.08
26	423	-2.28	0.08
27	395	-1.76	0.08
28	380	-1.53	0.08
29	395	-1.76	0.08
30	410	-2.02	0.08
31	379	-1.52	0.08
32	395	-1.67	0.08
33	389	-1.67	0.08

SAMPLE 11 (N=923)			
ITEM	ITEM SCORE	EASI-NESS	STD. ERROR
1	576	-0.62	0.08
2	75	2.96	0.15
3	112	2.41	0.12
4	225	1.38	0.09
5	156	1.94	0.11
6	296	0.91	0.09
7	250	1.21	0.09
8	304	0.87	0.09
9	208	1.50	0.10
10	213	1.47	0.10
11	290	0.95	0.09
12	169	1.82	0.10
13	223	1.39	0.09
14	270	1.08	0.09
15	250	1.21	0.09
16	155	1.95	0.11
17	136	2.14	0.11
18	203	1.54	0.10
19	122	2.29	0.12
20	153	1.97	0.11
21	133	2.71	0.11
22	150	2.00	0.11
23	242	1.26	0.09
24	376	0.46	0.08
25	181	1.72	0.10
26	370	0.49	0.08
27	108	2.46	0.12
28	235	1.31	0.09
29	245	1.24	0.09
30	173	1.73	0.10
31	155	1.95	0.11
32	162	1.89	0.11
33	217	1.44	0.09

SAMPLE 1 (N=488)			
ITEM	ITEM SCORE	EASI-NESS	STD ERROR
34	376	-1.48	0.08
35	383	-1.58	0.08
36	315	-0.73	0.08
37	354	-1.18	0.08
38	365	-1.33	0.08
39	390	-1.68	0.08
40	395	-1.76	0.08
41	394	-1.75	0.08
42	396	-1.78	0.08
43	383	-1.58	0.08
44	395	-1.76	0.08
45	369	-1.38	0.08
46	368	-1.37	0.08
47	388	-1.65	0.08
48	372	-1.42	0.08
49	407	-1.97	0.08
50	399	-1.83	0.08
51	356	-1.21	0.08
52	387	-1.64	0.08
53	388	-1.65	0.08
54	375	-1.46	0.08
55	374	-1.45	0.08
56	347	-1.10	0.08
57	390	-1.68	0.08
58	414	-2.10	0.08
59	406	-1.95	0.08
60	404	-1.91	0.08
MEAN	380.68	-1.57	0.08
STD. DEV.	22.87	0.34	0.00

SAMPLE 11 (N=923)			
ITEM	ITEM SCORE	EASI-NESS	STD. ERROR
34	102	2.54	0.13
35	255	1.17	0.09
36	116	2.34	0.12
37	169	1.82	0.10
38	233	1.32	0.09
39	182	1.71	0.10
40	159	1.91	0.11
41	150	2.00	0.11
42	139	2.11	0.11
43	124	2.27	0.12
44	180	1.73	0.10
45	154	1.96	0.11
46	100	2.57	0.13
47	155	1.95	0.11
48	156	1.94	0.11
49	193	1.62	0.10
50	150	2.00	0.11
51	119	2.33	0.12
52	147	2.03	0.11
53	137	2.13	0.11
54	136	2.14	0.11
55	77	2.92	0.15
56	161	1.89	0.11
57	109	2.45	0.12
58	147	2.03	0.11
59	162	1.89	0.11
60	177	1.75	0.10
MEAN	187.07	1.75	0.10
STD. DEV.	82.16	0.61	0.01

The results presented in table 12 show easiness values obtained from two contrasting ability groups, designated high and low 9sample 1 and sample 2 respectively). For sample 1, easiness values ranged from 2.57 assigned item one, a very easy item, to an index of -0.73, assigned item 36 (x = -1.57, SD = 0.34). For sample 2, easiness values ranged from -0.62 assigned item one an exceedingly easy item to an index of 2.92 assigned item 55, a substantially difficult item (x =1.75,SD =0.61). To perform this analysis easiness values from two samples were correlated using Pearson product moment correlation and co-efficient value of 0.24 resulted. Details are presented in Table 13.

Table 13

Computer Analysed Data On Inter-Correlation Between Easiness Values Computed In two Samples

	Sample 1	Sample 2
Sample 1	1.00	
Sample 2	0.24	1.00
Means (X)	-1.75	1.57
Standard Deviation (SD)	0.34	0.61

The significance of this coefficient was evaluated using one-way analysis o variance, ANOVA (Leabo, 1972). The results of the one-way analysis of variance ar(presented in Table 14.

Table 14

ANOVA SOURCE TABLE FOR TESTING THE SIGNIFICANCE CORRELATION COEFFICIENT BETWEEN EASINESS ESTIMATES (N=60)

Source	df	Ss	Ms	F-ratio	Significance F
Regression	1	1.28	1.28	3.56	0.06
Residual	58	20.80	0.36		
Total	59	22.08			

With the F - ratio as a test of significance, the linear regression co-efficient was not significant different from zero, $F = 3.56$ (df = 1, 58; n = 60; $P<0.05$). The null hypothesis of no significant relationship was not rejected based on available statistical evidence. This means that there is no significant relationship between the latent trait model item difficulty indices obtained from two contrasting ability, groups of the same Standardisation sample. The non-significant relationship implies that the difficult estimates are not invariant across groups of examinees that differ in ability.

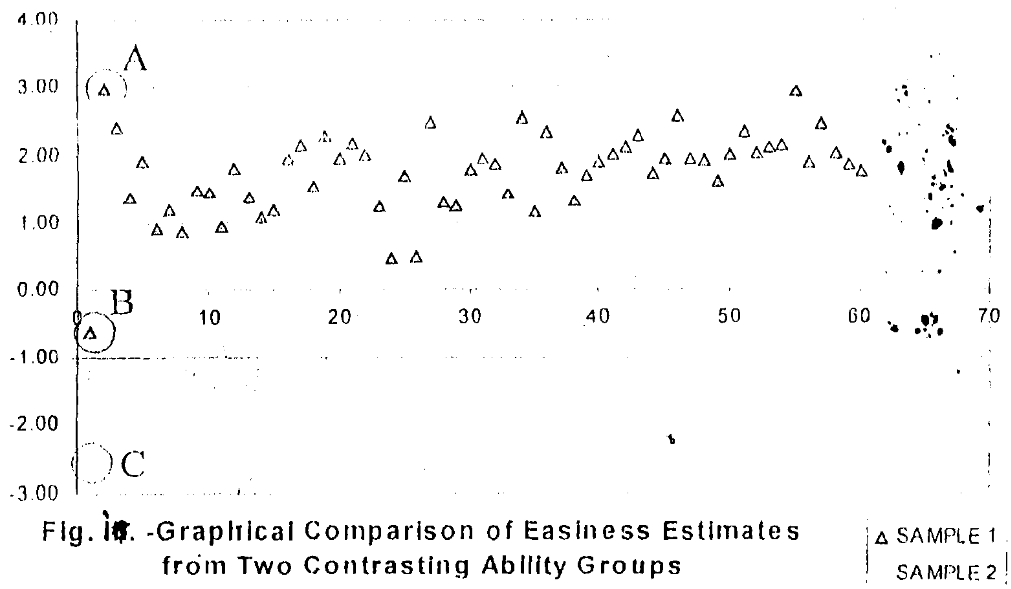


Fig. 1?. -Graphical Comparison of Easiness Estimates from Two Contrasting Ability Groups

Figure I shows graphical comparisons of easiness estimates from two contrasting groups.

Figure I indicates that the high ability group (sample 1) shows a slightly more dispersed distribution than the low ability group (sample 2). Three data points, representing three items, have been circled on figure I. The item identified, as "A" is a difficult item. The item identified as "C" is an exceedingly easy item. The item identified as "B" represents a different problem. The item appeared to be substantial more difficult for sample 2 (low ability group). Such an item is functioning differently in the two-ability group and may indicate item bias or an aberrant item. What ever the reason, the assumed invariance is missing. The item would be discarded regardless of any conventional item statistics suggesting its worth.

Hypothesis Two

There will be no significant relationship between ability estimates obtained from two contrasting score groups of the Standardisation sample ($P<0.05$).

The Pearson Product - Moment Correlation was used to test this hypothesis. Details are presented in Tables 15 and 16.

Table 15

Ability Estimates Obtained From Two Contrasting Score Groups

SAMPLE 1 (N=488)			
ITEM	**ITEM SCORE**	**EASI-NESS**	**STD ERROR**
40	16	0.70	0.23
41	32	0.78	0.23
42	32	0.86	0.24
43	28	0.94	0.25
44	42	1.03	0.25
45	47	1.12	0.26
46	27	1.21	0.27
47	57	1.31	0.28
48	38	1.41	0.29
49	54	1.52	0.31
50	53	1.64	0.32
51	17	1.76	0.34
52	7	1.90	0.36
53	8	2.06	0.38
54	2	2.23	0.41
55	12	2.44	0.45
56	6	2.68	0.51
57	4	2.99	0.59
58	4	3.42	0.72
59	2	4.14	1.02
60	0		

SAMPLE 11 (N=923)			
ITEM	**ITEM SCORE**	**EASI-NESS**	**STD. ERROR**
1	4	-4.14	0.13
2	5	-3.42	0.13
3	5	-2.99	0.13
4	17	-2.68	0.14
5	16	-2.44	0.14
6	36	-2.23	0.14
7	43	-2.06	0.14
8	46	-1.90	0.14
9	74	-1.76	0.14
10	59	-1.64	0.14
11	103	-1.52	0.15
12	73	-1.41	0.15
13	80	-1.31	0.15
14	94	-1.21	0.15
15	50	-1.12	0.15
16	69	-1.03	0.15
17	37	-0.94	0.15
18	70	-0.86	0.16
19	42	-0.78	0.16
			60

TOTAL COUNT	488
MEAN OF ABILITY TEST	1.35
STD DEV. OF ABILITY TEST	0.50
MEAN OFSTD. ERROR	0.29
STD. DEV. OF STD. ERROR	0.07

TOTAL COUNT	923
MEAN OF ABILITY	-1.46
STD. DEV. OF ABILITY	0.51
MEAN OF STD. ERROR	0.15
STD. DEV. OF STD. ERROR	0.01

The results presented in Table 15 show a comparison of ability estimates assigned to a given raw score by two samples of different abilities. For sample 1 (high ability group), the ability estimates ranged from 0.70 to 4 14 (x = 1.35, SD = 0.50). For sample 2 (low scoring group), the ability estimates ranged from - 4.14 to -0.78 (x = - 1. 46, SD = 0.51). To determine the relationship between the two ability estimates, correlational analysis was performed and a co-efficient value of 0.87 obtained. Details are presented in Table 16.

Table 16

Computer Analysed data on Inter-correlation Between Ability Estimates obtained from Two Contrasting Ability groups

	Sample 1	Sample 2
Sample 1	1.00	
Sample 2	0.87	1.00
	x = 1.35 SD = 0.50	x = -1.46 SD = 0.51

The significance of this correlation co-efficient was tested using the -ratio. The results of one-way Analysis of Variance, ANOVA, are presented in Table 17.

Table 17

Analysis of Variance for the sample Regression of Ability Estimate of sample 1 and sample 2

Source	df	Ss	Ms	F-ratio	Significance F
Regression	1	8.30	8.30	51.98	0.00
Residual	17	2.72	0.16		
Total	18	11.02			

With the F-ratio as a test of significance, the linear correlation co-efficient was significantly different from zero , F (1,17 = 51.98; n = 18; P<0.00. The correlation is highly significant from the standpoint of statistical sampling. The null hypothesis of no significant relationship is therefore, rejected. The significant relationship implies that the ability estimates are similar as those of the subjects in the standardisation sample. This implies that the means of the population from which the two samples were selected are equal. Figure ii illustrates further the relationship between ability estimates computed from the raw scores of 488 "high ability" and 932 "low ability" subjects.

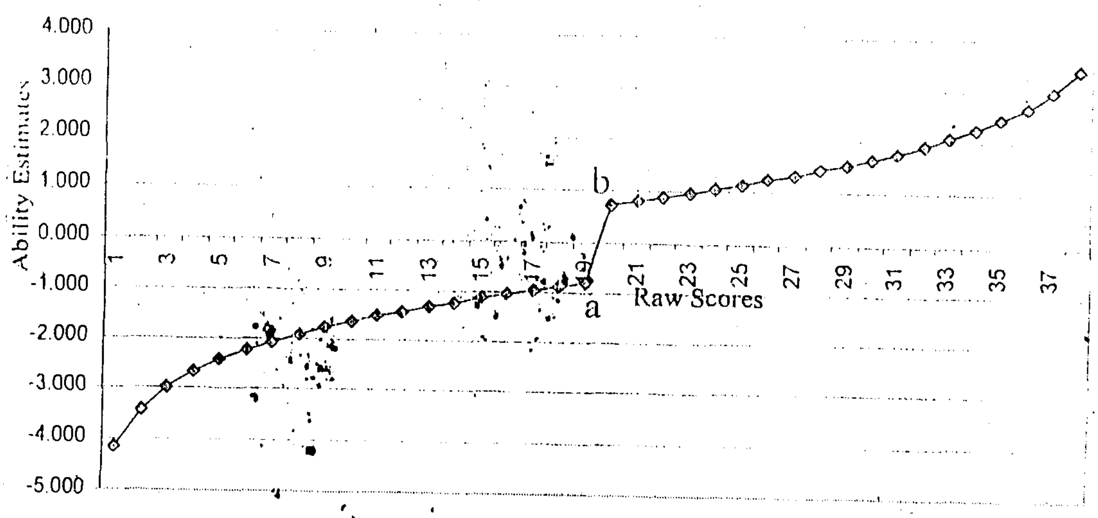


Fig ii Invariance of Latent Trait Ability Estimates

It is apparent from figure ii that the two ability estimates are practically identical and almost constant. This means that there is no systematic variance in the ability estimates. The sharp cut line, a to b, is apparently due to sampling fluctuation occasioned by the elimination of the middle ability group, numbering 804 subjects from the standardisation process.

Hypothesis Three

There will be no significant relationship among the item parameters obtained from the one—parameter model, the two-parameter model and the classical point biserial correlation co-efficient ($P<0.05$).

To test this hypothesis, item validity co-efficients computed from each model, that is, the fit statistics 9one-parameter model), the discrimination indices (two-parameter model), and the point biserial (classical test model) were intercorrelated. The results are presented in Table 18. The means and standard deviation are listed in the column below each variable (model).

Table 18

Matrix of Intercorrelations Among Three Methods of Estimating item Validity Indices (n=60)

Variable	One-parameter	Two-parameter	Point biserial
One-parameter	1.00		
Two-parameter	0.50	1.00	
Point Biserial	0.67	0.81	1.00
Mean	1.26	0.98	0.48
Standard Deviation	0.9	0.25	0.08

The correlation among the validity indices is high and positive. In terms of absolute magnitude, the co-efficients are relatively uniform, being 0.50 for pair one-parameter and two-parameter models 0.67 for pair one-parameter and point biserial, and 0.81 for pair two-parameter and point biserail. The high and positive correlation co-efficients imply that the validity estimates are essentially the same. The most conspicuous similarity was the relationship between the two-parameter and the classical point biserial. In order to determine if a statistically significant relationship existed among these co-efficients, a two-way analysis of variance (ANOVA) was performed 9Leabo, 1972). The results are presented in Table 19.

Table 19

F-ratio Summary Data of a 3 x 2 Multivariate Analysis of Variance (MANOVA)

Source of Variation	Ss	Df	Ms	F-Value
Rows (Validity indices)	3.83	59	0.065	3.61*
Columns (Methods)	18.56	2	9.280	515.56**
Error	2.10	118	0.018	
Total	24.49	179		

* Significant at P<0.05

** Significant at P<0.00

Results of the mutivariable analysis of variable show that the validity indices (between rows) are highly significant, F = 3.61 (df = 59, 118; n=60;P<0.05). This means that the validity indices are substantially similar.

Again, based on the degree of freedom of 2 and 118 for the between columns (models), the classical and latent trait models are highly significant (F=515.60,n=60'P<0.00). This means that the classical and latent trait models provides essentially similar item statistics, although the latent trait model analysis provides effective and precise parameter estimates as compared to the classical test model. The null hypothesis of no significant relationship is thus rejected. This implies that there is a significant relationship among the item parameters obtained from the one-parameter, and the two-parameter latent trait models and the classical point biserial correlation.

After the analysis comparing the item statistics computed from the one-parameter, the two-parameter and the classical point biserial was done, a more global analysis was performed. This analysis compared the value obtained from the one-parameter, the two-parameter model, the classical item analysis and factor analysis to determine a general relationship and to investigate if fit is related to factorial validity. A combined correlationed matrix indicating the relationship among all variables concerned in the comparison are presented in Table 20.

Table 20
Correlations Among One-Parameter Model, Two-Parameter model Classical Point Biserial and Factor Loading

Variable	One-parameter	Two-parameter	Point biserial	Factor Loading
One-parameter	1.00			
Two-parameter	0.50	1.00		
Point Biserial	0.67	0.81	1.00	
Factor Loading	0.49	0.97	0.79	1.00
Mean	1.26	0.98	0.48	0.49
Standard Deviation	0.19	0.25	0.08	0.09

Results show that the two-parameter, discrimination parameter estimates correlate highly with the first factor loading. The relationship was almost perfect. The fit statistics from the one-parameter latent trait model correlated moderately with the first factor loading. This reflects the fact that the two-parameter model measured the first factor of the physics achievement test, while the fit statistic measured a combination of all the factor.

Hypothesis Four

There will be no significant relationship between the classical raw scores and their logit exponential transformation ($P<0.005$).

Another comparison made between the classical test model and the latent trait model was concerned with the unit used to express ability. In the classical test model, examinees abilities are expressed as number of items correctly answered. The latent trait model analysis, on the other hand, transformed this value into a logarithmic function. Ability estimate from the latent trait model was compared to raw scores obtained from the classical test model to examine.

1. To what extent do the two models provide similar ranking of examiners, and
2. How close were the scores to each other? The correlations between ability estimates based on the two models and a sample of 2215 subjects are summarised in Table 21.

Table 21

Correlation Between Ability Estimates (N =2215)

	Raw Scores	**Ability estimate**
Raw Scores	1	
Ability estimate	0.99	1
Means (x)	25.97	0.33
Standard Deviation (SD)	14.19	1.18

The correlation between ability estimates was nearly perfect. This means that ability estimates based on the latent trait models and classical test model are, to a large extent, dependent on the number of correct answers. Given the observed nearly perfect correlation co-efficient between the two ability estimates, the next aspect investigated was the agreement between the scores, that is, how close the scores were to each other. This comparison was done using the mean absolute difference(Glass & Hopkins, 1984).

The t-score transformation (Glass & Hopkins, 1984) was applied to the ability estimates obtained from each analysis in order to express the ability estimates in the same unit (mean = 50, standard deviation = 10). The mean absolute difference between the transformed scores were computed using (Ndalichako & Rogers, 1997)

$$MAD_{LC} = \sum_{j=1}^{N} \frac{(T_L - T_C)}{N}$$

ıere MAD_{LC} is the mean absolute difference between the ability estimates based on tent trait model (L) and the classical test model (LC), T_L and T_O are the corresponding ansformed scores, and N =2215 (number of subjects).

ransformed raw score = $\frac{\text{Subject's raw score - mean of the test x 10 + 50}}{\text{Standard deviation of test}}$

$\sum T_C = 110750.00$

ansformed ability estimate = $\frac{\text{Subject's ability - mean ability x 10 + 50}}{\text{Standard deviation of ability}}$

$\sum T_L = 110750.00$

Mean Absolute Difference (MAD_{LC}) = $\frac{110750.00 - 110750.00}{2215}$

$= 0.00$

The mean absolute difference between the two pairs of transformed scores was zero. This means that the agreement between the classical raw score and the log ability estimate is perfect. The null hypothesis of no significant relationship ($P<0.05$) is therefore not rejected based on statistical evidence

Summary of Result

The following results were obtained

1. A factor analysis of the physics achievement test resulted in 14 Eigen values greater than 1.0, the first Eigen value, 14811 accounted for 25.104 percent of the variances.
2. The physics achievement test item parameters, difficulty indices discriminating indices and the associated standard errors were reasonably estimated
3. The amount of ability necessary to achieve a given raw score was reasonably estimated.

2. The physics achievement test item parameters, difficulty indices discriminating indices and the associated standard errors were reasonably estimated
3. The amount of ability necessary to achieve a given raw score was reasonably estimated.
4. An analysis of model fit shows that two items failed to fit the model
5. The overall reliability co-efficient (KR-20) of the physics achievement test was 0.89.
6. The inter-correlations among the various content areas of the physics achievement test were high and positive. The inter-correlations were in the range,0.58 to 0.73.
7. There was no significant relationship, statistically, between easiness estimates obtained from two contrasting score groups of the same Standardisation sample
8. There was a significant relationship between ability estimates obtained from two contrasting score groups of the same Standardisation sample.
9. The inter-correlation among validity indices obtained from the one-parameter model, the classical point biserial and factor loading was high and statistically significant.
10. The correlation between ability estimates based on the ability estimates are identical.

CHAPTER FIVE

Discussion Of Results, Conclusion, Implications and Recommendations

In this chapter, an attempt will be made to conceptualise and generalise the findings of the study in relation to the specific objectives and the results of previous studies. The discussions are presented under each research question or hypothesis formulated. Following the discussion are conclusions, implications of the study, recommendations, limitations of the study, suggestions for further research and summary of the study.

Discussion of Research Question One.

> The purpose of this research question was to investigate the unidimensionality, single trait for ability of the physics achievement test.

A factor analysis based on varimax rotation (Orthogonal Option) was used to investigate the unidimensionality of the final version of the physics achievement test. The analysis resulted in 14 Eigen values greater than 1.0. The first Eigen value, 14.811 accounted for 25.104 percent of variance. This satisfied the minimum criterion for establishing the unidimensionality of the physics achievement test (Reckase, 1979). According to Reckase (1979), the existence of a single factor would also imply unidimensionality. The findings of the study satisfied Reckase's recommendations. The results of the study are also consistent with the findings of Reiser (1983); Harrison (1986), Goldman & Raju, (1979). In a discussion of model's misfit, Harrison (1986) presents three factors in test performance that would violate unidimensionality. These are: Instructional emphasis, speed of work and individual propensity to guess. According to Harrison, differences in content and differential instruction in the target population may contribute to lack of unidimensionality for a typical achievement test. These factors could make undimensional latent trait models untenable for achievement tests.

Lord (1980), advocates three reasons for a statistical test for unidimensionality.

1. It is often vital that a test that purports to measure the level of a certain ability is indeed not significantly contaminated by varying levels of one or more other abilities displayed by the examinees taking the test. For instance, a test of physics achievement should not be confounded by varying levels of verbal ability displayed by the examinees.
2. It is essential that a test designed to be used in the measurement of individual differences must in fact measure a unified trait.
3. Unidimensionality must be satisfied if much of the latent trait model methodology as incorporated in LOGIST is to be valid.

Although the factor analytic procedure has been used to establish unidimensionality, there are two problems associated with such a factor analytic procedure. According to Hambleton, et-al (1978), the first problem associate with this procedure is that the matrix of intercorrelation may not be positively definite and hence in the strict sense cannot be factor-analysed. The second problem identified is that the latent trait models are non-linear, whereas factor model is linear.

Discussion of Research Question Two

> The purpose of this research question was to estimate the item parameter, latent trait item difficulties using the one-parameter model.

Results of research question two show that the items ranged in difficulty from - 1..49 to +0.49, suggesting that the physics achievement test covers a wide spectrum of ability of senior secondary school physics students. The results support previous findings (Korashy, 1995; Harrison, 1986; Bryce, 1981) and the results of pilot study. According to Harrison (1986), item difficulties lower than -2.5, and higher than +2.5 are usually not useful for achievement testing. The difficulty estimates, therefore, fall within the recommended range. The distribution of the item difficulty estimates was roughly normal, having a mean of zero and a standard deviation of 0.31. The results are consistent with the findings of pilot study and the study by Allen, et-al (1987).

Differential instruction and content coverage are factors which affect item difficulty estimates (Allen et-al 1987). According to Allen, et-al, differential instruction may affect difficulty estimates to such an extent that the test may no longer be considered unidimensional for the standardisation sample. To improve estimation, Harrison (1986) recommends that test developers should increase the number of common factors within a test. Creating various item formats, for instance, sentence completion items and matching items within a test may be one way to comply with these recommendations. These recommendations for improving the estimability of item parameters from a set of test items are very similar to Bryce (1981) suggestions for building systematically heterogeneous test.

Discussion of Research Question Three.

> The purpose of this research question was to estimate the item parameters, difficulty and discriminating indices, using the two-parameter latent trait model.

The results of the two parameter latent trait model item analysis show that the estimated discrimination indices ranged in value from 1.76 to 0.39. Results further indicate that the more difficult items and items that fit the latent trait model discriminate substantially amongst the subjects because of discrepant factors on either the teaching of the relevance topics or in the formulation of the test itself. The results agree with the findings of previous researchers (Example: Lord, 1968; Bryce, 1981, and Harrison, 1986). One other factor responsible for the high discrimination indices in the present study is that the items are highly interrelated. Urry (1974), notes that the discrimination parameter in the latent trait model will be systematically under estimated and the difficult indices over estimated because of error variance, (unreliability) in ability estimates. According to Urry (1974), discrimination estimates also depend on such factors as sample size, test length, and test reliability. According to Urry (1974), these measures must be sufficiently large to effectively reduce the spurious component of part-whole correlation that tends to attenuate discrimination estimates. Another problem that remains with the discrimination and difficulty estimates is that its preciseness depends upon the underlying normality assumptions (Lord,1968).

Discussion of Research Question Four

> The purpose of this research question was to estimate the subject measures, ability estimate and its associated standard errors.

Validating a physics achievement test using the latent trait psychometric procedure results in logarithmic ability estimates being assigned to a given raw score ranging from - 4.0 to + 4.0 (Ludlow & Haley, 1995). The results show that the latent trait ability estimates produced a substantially linear section in the region -4.14 to 4.14 on the log scale and fall within the recommended range. The results are consistent with the findings of previous investigators (Tinsey & Dawis, 1995, Douglas Khavari & Farber, 1979; Gallini,1983; Harrison,1986). Ludlow and Haley (1995:969) give the following opinion about the estimate of a person's ability.

"A person ability estimate in logits is their (his or her) natural log odds for succeeding on items of the kind chosen to define the zero point on the scale. Logits are defined mathematically as the natural log of an odd ratio". According to Cox (1970),odds ratio tells the relative frequency of an event occurring versus the relative frequency of it not occurring . For example, to say that the odds of an event occurring in one's favour are 2/1 means that out of 3 times, the event is expected to occur twice in one's favour . The natural log of 2/1 is 0.69 (a positive logit). If the odds of the event occurring in one's favour are 1/2, then the natural log is -0.69 (a negative logit) . Finally, if the odds are 1/1, the natural log is o. This means that frequencies of an event occurring in testing situation translate into the number of items accomplished or failed. Suppose there are 60 PAT items . If one succeeds or passed 40 and failed 20 of them, one's odd ratio for the test is 40/20 or 2/1, and one's ability estimate as a logit is the natural log of 2 which is 0.69. If one succeeds on half the items (odds of success are 1/1, one's ability logit is zero and If one fails on 40 items (odds of success are 1/2), one's logit estimate is -0.69. Thus any ratio of the number of successful passes to failures can be converted to a logit (except for zero successes or zero failures) because the natural log of zero is undefined mathematically.

Smith (1985), identifies measurement disturbances such as guessing, start-up, content/person interaction and sloppiness as sources of potential error in the estimation of ability estimates, being either over or under-estimated. The first type of measurement disturbance identified is guessing. Random guessing on multiple choice item is a particular problem because not all examinees exhibit the same guessing capacity. Some guess frequently whether they are instructed to or not, while the more conversative will rarely guess. Propensity to guess at random may lead to over estimation of a subject is true ability. Some authors such as Allen & Wendy (1979), and Harrison (1986) suggest the use of the three-parameter latent trait model as a solution to guessing.

The second type of measurement disturbance identified is start-up. Smith (1985 : 440) gives the following opinion about start-up: "Start-up is a term used to describe the situation that occurs when a person who has a high test anxiety for a test performs poorly on the items at the beginning of the test, but after over coming his anxiety, goes on to perform better. Due to this anxiety, the person appears less able on the initial items and more able on the later items".

Thus, the ability estimate based on the entire answer to the items will under estimate his ability. As a solution to this problem, Smith (1985) suggests the use of robust estimation procedure. Robust estimation is a statistical procedure that is designed to lessen the impact of unexpected answer patterns on the estimation of the person's ability. This procedure is based on the calculation of the probability of a correct answer to an item/person interaction within the framework of a latent trait model (Smith, 1996).

The third type of measurement disturbance identified is content/person interaction. According to Smith (1985), content/person interaction is a condition where a person is unfamiliar with one or more of the content areas, due perhaps to non-uniform curriculum, instructional emphasis and content coverage. To lessen this effect on the ability estimate, Smith (1985) suggests another robust estimation procedure. This analysis seeks to identify the cause and magnitude of the measurement disturbance and is referred to as person analysis. That is, the analysis of a person fit to the model. Person analysis is readily accomplished using the unweighted total fit statistics (Smith, 1996 : 1985).

The fourth type of measurement disturbance identified is sloppiness, carelessness. This is, lack of concentration resulting in carelessly reading the questions. Examinees sometimes failed to correctly match up distractors with marks on the answer sheet. Birenbaun (1985), mentions tendencies such as sleeping, fumbling and plodding as factors which cause error in the estimation of true ability estimates. According to this terminology (Birenbaun, 1985), sleepers are examinees who get bored with the test and do poorly on later items, while fumblers are the ones who do poorly in the beginning because of confusion with test format. The ones who never get to the later items on the test are the plodders. As latent trait theory becomes more complete, the role of ability estimates in achievement testing will become more prominent. Meanwhile, ability estimates are based upon prior item easiness and PROX techniques as exemplified in the work of Wright & Panchapakesan (1969) and Izard & White (1980) respectively.

The standard error associated with ability estimates is a measure of the precision attained in estimating the ability. The mean of the standard error associated with the ability estimates is 0.01 of the standard deviation value of the ability estimates. The results support previous findings (Gallini, 1983, Korashy, 1995). The standard error associated with ability estimates depends primarily upon the number of items composed of the test and the range of easiness values (Wright & Panchapakesan, 1969). This means that the item easiness should be reasonably appropriate to the abilities being measured by the test.

Discussion of Research Question Five

> The purpose of this research question was to estimate standard error for each of the PAT items.

In latent trait analysis, each item difficulty index or ability estimate is accompanied by its standard error, the most rudimentary indicator of reliability in the classical test model. Inferences and validations on an item utilising the latent trait model are therefore derived from a solid statistical basis. Table 5 shows that the standard error estimates ranged from 0.0578 to 0.0518. This reflects a good precision of the test, and supports Korashy (1995), Gallini (1983) findings. Harrison (1986) obtained similar results and suggested that test developers should increase the number of relevant test items in order to improve the accuracy of estimation. The smaller the standard error values, the narrower the confidence bands around the trait being measured (Bock, 1972). Thus for each item composed of the test, it is possible to obtain more precise confidence statements than one index of reliability of the total test computed in the classical test model. According to Bock (1972), the estimated standard errors become exact, that is zero, as the sample size increases. This means that sample size is one factor that influences the standard error, reliability of a test item.

It is meaningless to talk about standard error without specifying the sources of measurement error that are being considered. One important source of measurement error is the selection of specific items to test a general ability (Livingston, 1988). It results from the fact that a test score based on a sample of items will be only an approximate indicator of the subject's ability. Livingston (1988), suggests three ways to determine the influence of this source of measurement error. One way is to give the same subject two different versions of the test. These different versions of the same test are called alternate forms, and this type of reliability is called alternate forms reliability. Another way is to split the test into two similar half tests. This type of reliability is called split-half reliability. According to Livingston (1988), a third way to determine the influence of this reliability is to analyse the data from the individual test items. This type of reliability is called internal consistency.

A second important source of measurement error is the time of testing. According to Livingston, this source of error include short-time memory, and temporary changes in the condition of testing. To determine the influence of this source of measurement error, Livingston (1988) suggests 'that the subject be given the test two or more different times. The agreement in the test results for the same subject at different time is called stability. A third important source of measurement error suggested by Livingston is the person who grades or scores the test. To determine the influence of this source of measurement error, it is suggested that two or more graders score the same test paper. The agreement between the results obtained from different graders is called inter-rater reliability.

Discussion of Research Question Six

> The purpose of this research question was to investigate the fit of the physics achievement test items to the one-parameter latent trait model.

The analysis of the model's fit shows that two items (Nos. 36 and 46) failed to fit the model as evidenced by the large positive fit statistic. In an attempt to explain the misfits, it was noted that items 36 and 46 were extremely difficult for the subjects, (difficulty values, b_{36} = 0.49, b_{46} = 0.47). On further examination, it was noted that item 36 was an item testing a higher level complexity dimension (synthesis). It is possible that the item was difficult for the subjects because of the higher level cognitive processes involved in answering the item. Item 46 was an item on the lowest level of the cognitive domain (knowledge). As Gallini (1983) notes oftentimes extremely easy and difficult items result in misfit because of careless mistakes and considerable guessing among the examinees. Both occurrences contribute to discrepancies in the proportions expected to answer the item correctly in the different ability groups. A Comparison between the observed proportions and those expected to correctly answer item 36 is desirable. Table 8 shows that more subjects in the low ability group answered item 36 correctly (13 percent Vs 0.2) probably due to "lucky" guessing . This means that higher ability subjects failed the item.

According to Harrison (1986), the failure of an item to fit the model can be attributed to two main sources. One is that the model is too simple . It takes account of only one item parameter, item easiness. The other source of lack of fit lies in the content of the items. The model assumes that the items are measuring the same trait. Thus items in a test may not fit the model, if the items are so badly written that what is measured is irrelevant to other items composed of the test. Lord (1968) points out that individual items which discriminate well among the subjects fit latent trait model rather well.

Discussion of Research Question Seven

> The purpose of this research question was to estimate the classical test model item parameter, point biserial correlation co-efficients in order to provide a baseline against which latent trait model item parameters will be compared.

A further analysis performed on the physics achievement data was the classical test item analysis. The purpose was to obtain classical item statistics as a basis for comparison with the latent trait model. The results of the classical test model item analysis show that a majority of the items have relatively low difficulty values, because some schools have not covered sufficient contents as specified in the Senior School Certificate syllabus. The findings are consistent with Nworgu's (1985) findings . The high point biserial values can be explained in terms of the effectiveness of the item review exercise following the pilot testing and critical comments by experts. One other possible cause for the high point biserial correlation is that the item has good construct validity, that is the empirical factor structure of the physics achievement test corresponds to the theoretical Organisation of the trait measured by the physics achievement test (Bentler,1980).

Other factors suggested for the high point biserial correlations among test items are representativeness of content (content validity) and interrelatedness among the items (Ebel, 1968). This means that a high point biserial correlation co-efficient is synonymous with the extent to which the test items covers a representative sample of the behaviour domain to be measured. Bryce (1981), noted that reliance on high item point biserial Correlation co-efficient alone might lead to the acceptance or rejection of an item for the right or wrong reasons. According to Bryce, highly discriminating items may reflect variation in more than one kind of ability in the sample. Bryce opines that to the extent that complexity level reflects different kinds of ability, the test developer should also ensure heterogeneity of content.

One important purpose of the classical test model is to minimise error variation in the total score computed from a measurement instrument (Lulow & Bell, 1996). The principal tools used in this study for investigating error variation were reliability and validity of the physics achievement test. Specifically, Kuder and Richardson (1937), formula 20 was used to assess the physics achievement test's internal consistency and factor analysis was used to examine its construct validity. The reliability estimate for the physics achievement test was high (0.89), implying that the physics achievement test is valid and reliable. The reasons for this reliability estimate can be attributed to the number of items composed of the test which were sufficiently large and the superior item – total test correlation co-efficient.

Discussion of hypothesis one

> The purpose of this hypothesis was to investigate the invariance of the latent trait item difficulties with respect to the ability of the standardising sample.

The first of the problems identified with the classical test model was that item statistics, easiness values depended solely on the distribution of the ability of subjects in the sample used in the validation process. The items easiness is therefore not invariant across a group of subjects that differ in ability. It was hypothesised that in the latent trait model the item easiness can be computed independently of the nature of distribution of abilities in the research sample. The correlation between difficulty indices computed in sample 1 (high ability), and those in sample 2 (low ability) was 0.24. The correlation co-efficient of 0.24 was positive but not statistically significant at 0.05 alpha level. This implies that item difficulty indices are not invariant across a group of subjects that differ in ability. Previous investigators (example; Wright, 1967); Anderson et al, 1968; Tinsley & Dawis, 1975 and Guiton & Ironson, 1983) have reported a similar low, but positive correlation.

One factor which reduced the invariance of the item easiness values in the present study was that subjects with low scores were included in the estimation process. Wright & Panchapakessan (1969), suggests that subjects with extreme low scores should be eliminated from the sample so that the validation process is not contaminated by guessing. It is possible, therefore, that guessing may have reduced the invariance of the item easiness in the two standardisation ability groups. Hambleton and Cook (1977), notes that the computed easiness values, will in practice, differ from one sample to another if the samples differ in mean value or variability, on the underlying trait. Hambleton and Cook (1977), further note that item easiness is invariant when estimated on samples that differ in ability but only within a linear transformation, that is, the property of items easiness invariance is a consequence of specifying the model, so that the logit of the probability of answering an item correctly is a linear function of the latent variable.

Another factor which reduced the item easiness invariance was that mis-fitting items were included in the comparison. As Tinsely & Dawis (1975) point out, invariance hypothesis cannot be expected to hold for items that do not fit the model. Differential instruction and content coverage are other factors that reduced the invariance easiness hypothesis (Allen, et al, 1987). According to Allen et al, some subjects in the standardisation sample may have come from schools where the emphasis on instruction is different from that of other schools. Thus, if the item easiness has been estimated using latent trait procedures on the basis of two standardisation samples in which differential instruction has occurred, the item easiness may not be invariant (Allen, et al, 1987). The concept of invariance difficulty indices across groups of examinees is also a function of the physics achievement test meeting the assumptions of the latent trait model, such as; single trait for ability of the physics achievement test and local independence among the physics achievement test items (Baker, 1977; Allen & Wendy, 1979).

Discussion of hypothesis two:

> The purpose of this hypothesis was to investigate the invariance of the ability estimates with respect to the ability of the standardising sample.

A comparison of the ability estimates assigned to a given raw score by two samples of different ability indicates that the ability estimates are invariant with respect to ability of subjects in the standardisation sample. The correlation between ability estimates computed in sample 1 (high ability) and those in sample 2 (low ability) was 0.87. The correlation co-efficient of 0.87 was positive and statistically significant at 0.05 alpha level. Previous researchers (example; Bright, 1967; Anderson et al; 1968; (Tinsley & Dawis, 1975) have compared ability parameter estimated from different ability groups formed randomly on the basis of the raw scores and obtained similar high, positive and statistically significant correlation co-efficients.

In the classical test model, reliability estimates are specifically a function of a particular set of items and sample of examinees on which the data have been collected (Weiss & Davison, 1981). This means that the same individual tested with two different samples or groups of examinees may obtained two different errors of measurement and estimates of true score based simply on the group of testees. The total raw scores are therefore not invariant with respect to ability of subjects in the standardisation sample (Weiss & Davison, 1981). In contrast to the classical test model, the standard errors computed for each item in the latent trait model analysis are independent of the particular examinees sampled or subset of items on which the data have been collected (Lord & Novick, 1968). Each items contributes to the standard error independently of all other items in the test. The ability estimates are therefore invariant across group of items or individuals in the standardisation sample.

However, the results are not as equivocal as they may appear. Anderson, et al, (1968) have pointed out that latent trait ability estimates do not tend itself to small samples. Generally, sample of 1000 subjects are needed to obtain stable ability estimates in the two ability groups (Goldman & Raju, 1986). The concept of invariance ability estimates across group of items may not hold for achievement tests measuring varying abilities (Allen et al, 1987). According to them, invariance ability estimates depend upon the similarity of content covered and upon similarity of instruction given to the group of examinees. Birenbaum (1985) identified unusual answer patterns on a test as another factor which may void the invariance ability estimates. Aberrant answer patterns can result from misconception concerning the subject matter, poor instruction, cultural bias, test anxiety, exceptional creativity or poor test taking strategies (Birenbaun, 1985).

Discussion of hypothesis three:

> The purpose of hypothesis three was to investigate the relationship among the one-parameter, the two-parameter and the classical item point-biserial correlation co-efficient.

Results presented in Table 19 show that both classical item analysis and latent trait analysis produce essentially valid and reliable physics achievement test which according to their respective theoretical assumptions represent the best set of items defining the underlying construct or latent trait. In the classical test analysis, item-total correlation, point-biserials were considered as the criterion (Table 9), while latent trait analysis was conceptualised by examining items displaying good fit, low chi-squares (Table 8).

Although, these two analyses resulted in several common items selected, there were considerable differences in terms of specific items rejected, relative fit of retained items and the total number of items selected. One notable difference was the fact that latent trait analysis rejected two items (Nos.36 and 46) and accepted 58 items. The classical test analysis, on the other hand, rejected only one item (No.1) and accepted 59 items as the best set of homogeneous items.

Thus, there is a considerable overlap between these two validation procedures. The results agree with the findings of Douglas et al, (1979), Ndalichako & Rogers (1997).

The common set of items accepted by both classical and latent trait procedures were items dwelling on translation, interpretation, extrapolation of specific elements in the physics achievement content (comprehension) and those of abstractions in particular and concrete situations or utility (Application). In addition, several items displayed excellent fit in one analysis but only marginal fit in the other. For example, item 2 (Table 9) yielded a high point biserial correlation (0.63) on the classical item and was therefore, a highly homogeneous item. However, in the latent trait analysis, the same item 2 (Table 8) yielded a marginal model fit (1.41). On the other hand, items 1 and 24 yielded good chi-square fit in the latent trait analysis, 0.49 and 0.82 respectively, but only marginal point biserial correlation co-efficients in the classical test model analysis. This means that although there were a common core of items selected, the classical and latent trait analysis procedures appear to yield different substrates of homogeneous items. Douglas, et al (1979) report similar findings.

A second comparison of the classical test model and latent trait model procedures was made in terms of correlation among the one-parameter model, the two-parameter model, the classical point biserial correlation and factor loading. In this analysis, the correlations with the factor loading are almost identical, leaving little basis for the choice of the two techniques in test validation, except cost. Some investigators (example, Reckase, 1979; Ludlow & Bell, 1996) have also reported similar findings.

Discussion of hypothesis four

The purpose of hypothesis four was to investigate the relationship between the classical raw scores and their logit exponential transformation.

Another comparison between the two analyses was concerned with the unit used to express ability. In the present study the log ability estimates were almost exactly proportional to raw scores, resulting in virtually no differences between initial raw score values and the logistic exponential transformation. The results support previous investigations (example; Ndalichako & Rogers, 1997; Ludlow & Bell, 1996; Douglas, et al.; Tinsley & Dawis, 1975). This could be explained by the fact that the ability estimates based on both the latent trait model and the classical test model are to a large extent, dependent on the number of correct answers. Thus, the examinees' positions are maintained despite the use of different models of scoring. Given the observed nearly perfect correlation between the two ability estimates, the next aspect investigated was the agreement between the scores. This comparison was done using the mean absolute difference (Glass and Hopkins. 1984). The mean absolute difference indicated a perfect agreement between the classical raw scores and their logit transformed scores. The findings agree with the results of Ndalichako and Rogers (1997).

One final practical consideration remains. On cost per analysis basis, the classical test analysis is clearly preferred. The classical test analysis appears to be more practical on the basis of its economy. In contrast to the complex latent trait analysis, classical item analysis requires less computer programme and much less time to analyze data set and is equally manually calculable. Thus, the classical item analysis can be executed with minimal of time and effort. Therefore, if meaningful differences continue to be elusive, cost efficiency and economy considerations must be considered as the primary practical distinction between the two procedures.

Conclusion

On the basis of the data analysed in this study, the following conclusions are drawn.

1. The physics achievement test is unidimensional, measuring one single trait for physics ability.
2. The physics achievement test covers a wide spectrum of ability of the senior secondary physics students.
3. The subjects generally responded in a consistent manner on the physics achievement test. This was seen in the acceptable range of the physics achievement test items fit to the latent trait model.
4. The physics achievement test is valid and reliable.
5. The latent trait model easiness estimates are not invariant with respect to the ability of the standardisation sample.
6. The latent trait model analysis leads to invariant ability estimates, while the classical test model analysis does not.
7. There is a considerable overlap among validity indices obtained from the one-parameter model, the two-parameter model and the classical point biseral correlation co-efficient. The two-parameter model measures the first factor of the physics achievement test.
8. The classical and latent trait models examined in the study provide essentially the same ranking of examinees. The examinees' position is maintained despite the use of the different models.

9. The latent trait model analysis provides effective and precise parameter estimates as compared to the classical test model. The classical test model however, appears useful in the first stage of test development in ensuring the effectiveness of distractors and is practical on the basis of its economy, and can be executed with minimum time and effort.

Implications of the study:

The findings of the study suggest the following implications:

1. The findings of the study might be addressed to those responsible for item banking (the West African Examinations Council, WAEC, the Joint Admission and Matriculation Board, JAMB). An item bank is a collection of test items "stored" with known item characteristics and made available to test developers (Hambleton, et al; 1978). Because of the invariant item parameters, known desired characteristics of the items can be stored and drawn from the bank when necessary.
2. The study has implications for latent trait investigation of item bias. Basically, the questions of bias ask whether an item is functioning differently in different groups. Guiton and Ironson (1983:75) reviewed some methods of item bias analysis and concluded that "due to its sample – invariant property, the latent trait model approach appears to be the only sound theoretical approach to the problem of item bias"
3. The study has implications for counselling since it permits the identification of each examinee's strength and weakness in testing situations. This could be achieved through analysis of person's fit to the model.

4. The study has implications in tailored measurement where the emphasis of testing is on obtaining different contents of items that can be used interchangeably without the classical requirement of a parallel forms test. Due to the invariant ability estimates, test items can be selected depending on the ability levels of each examinee.

5. The study has implications for criterion – referenced testing. Because of the unidimensionality of the PAT items, it is possible to use the test information function at each ability level to measure the power of an item to discriminate at a given cut-off between masters and non-masters. The information function of a test is the reciprocal of the square of the standard error associated with each item difficulty or ability estimate.

The study further has the following implications:

1. The findings of the study have implications for test designers. Since it is possible to get or obtain a precise measure of the standard error for each test item, it is possible to compare the relative efficiencies of different set of items for specific purposes.

2. The study has implications for test score equating. Because latent trait analysis provides difficulty and ability estimates on the same units, logits, the equating of the test forms can be done using either item difficulty or ability estimates.

3. The study has implications for test specialists in the detection of measurement disturbances such as guessing, startup, sloppiness and plodding through the analysis of each examinee's fit to the model.

Recommendations

On the basis of the findings of the study, the following recommendations are made.

1. For practical purposes, the latent trait analysis should be followed up by some qualitative analyses. This is particularly advisable as some item misfits cannot be determined by statistical results alone. Interview with examinees, for instance, can be conducted to ascertain the strategy they used to endorse a given item identified as the model misfit. Such observations can help identify and resolve contextual variables that contribute to poor model fit.
2. Further statistical treatment of items that showed a misfit in the latent trait analysis should be examined by the use of the nominal response model. The nominal response model provides additional and improved information in that it utilises information contained in the incorrect answers to the items (Book, 1972).
3. Since the study permits the identification of each examinee's strength and weakness in testing situations, this diagnosis should be used to improve the quality of physics instruction. Two major methods to improve the quality of physics instruction are recommended.
    a. Instructional Treatment. Physics teachers should develop instructional materials to fit the student's ability pattern. For instance, if a student is low in spatial ability, then a presentation emphasising the use of graphs, models and concrete demonstrations are suggested.
    b. Instruction with feedback, corrective procedures. Feedback devices for example formative evaluation, diagnostic tests should be built into

the instruction to identify deficiencies in the student learning of a given physics unit. The most corrective techniques include : re-teaching of a selected physics unit, small group study session and individualised tutoring. Essentially, the corrective devices will provide the students with the instructional cues, the learning participation and the reinforcements which are best suited to their characteristics and needs.

4. Before using latent trait analysis, it is highly recommended that substantive investigations be done to examine the latent structure of the test items, single trait for ability. If instructional emphasis, speed of work and propensities to guess create oblique common factors such as identified in the present study, then the three-parameter latent trait model that incorporates guessing parameter, is recommended.
5. Based on the findings of this study which show that the two – parameter model correlated highly with the factor analytic validation, future applications of the latent trait model might emphasize the two – parameter model.

Limitations of the Study

The generalisability of the findings of the study is subject to the following limitations :

1. The One-Parameter and the two-Parameter models on which the study was based are, in principle, applicable to data in which there are no omitted items. In this analysis, omitted items in the physics achievement test were treated as wrong answers. This introduced some inaccuracies in estimating

the ability parameter especially for subjects who did not have time to complete the test.

2. One factor which reduced the invariance of the item parameters and so limit the generalisability of the findings of the study is that subjects with low scores were included in the estimation of the item parameters of the physics achievement test. Wright and Pakan Pakesan (1969), suggested that subjects with low scores should be eliminated from the sample so that the validation process is not contaminated by guessing. This procedure was not carried out in the standardisation process.
3. In the statistical analysis carried out, post hoc decisions concerning the strategy used by each subject in answering the test items, were not carried out. This is particularly necessary as some item misfits could not be determined alone by statistical results.

The findings of the study are limited to a greater extent by problems involved in the collection of data. Because the physics contents have been taught differently by various teachers, the sample of the study included subjects from schools where the degree of emphasis on instruction varies from school to school. Furthermore, while some schools had already completed the senior secondary physics syllabus, in other schools, some content areas had not be mentioned prior to the collection of data. The differential instruction and dissimilarity in content coverage, which were practically unavoidable in the course of the investigation naturally attenuated the results.

Suggestions for Further Research

The researcher has attempted to present a thorough primer on how to carryout latent trait analysis. There are, however, a number of unresolved problems about latent trait analysis that need further research.

1. Further study should be carried out to investigate the model robustness, particularly when guessing does occur.
2. In order to extend the use of latent trait model across the full spectrum of mental testing, the applicability of the models across tests that are of different ability levels (e.g. Intelligence test) need further investigation.
3. Because the latent trait model does not hold for items which do not fit the model, the relationship between the invariance of the item parameters and the goodness of fit of the items need further investigation.
4. The use of the latent trait measurement model in the equating of a physics achievement test needs further investigation.
5. The use of the latent trait model in the development and analysis of attitudinal data needs to be further investigated.
6. Research should further be conducted on the effect of sample sizes on parameter estimation under multi-dimensional conditions. The present study used only one sample.
7. Finally, a latent trait validation of criterion referenced test needs to be researched further.

Summary of the Study

The development and validation of a physics achievement test has for long been founded on the classical test model. This is primarily due to its simplicity both conceptually and computationally. Unfortunately the classical test model has some serious disadvantages which make it no longer valid in establishing the psychometric properties of a physics achievement test. The limitations include: too restrictive and inadequate in scope to ensure objectivity for all items composed of a test. The major potential of the latent trait models rests in the characteristic of parameter invariance from which all other advantages emerged has several promising applications which make it worthwhile in considering latent trait analysis as a promising addition to the repertoire of measurement methods in the development and standardisation of a physics achievement test.

The study, therefore, was aimed at developing and establishing the psychometric properties of a physics achievement test using the one-and two parameter latent trait models on a sample of senior secondary physics students. The study also investigated the influence of the one-parameter and the two-parameter latent trait models and the classical test model on some psychometric properties of the physics achievement test. To accomplish this investigation, seven research questions and four hypotheses were formulated to guide the study. Specifically, the following research questions were investigated:

1. How unidimentional, single trait for ability are the physics achievement test items?
2. What are the estimates of the item parameter, difficulty indices using the one-parameter model?

3. What are the estimates of the item parameters, (difficulty and discriminating indices) using the two-parameter latent trait model?
4. What are the estimates of the subject measures, ability and its associated standard errors indicated by the raw scores on the physics achievement test (PAT)?
5. What are the estimates of the standard error, (reliability) for each of the PAT items?
6. To what extent do the PAT items fit the one-parameter latent trait model?
7. What are the estimates of the classical test model item parameters, difficulty indices and point biserial correlation co-efficient?

The following null hypotheses were formulated and tested at an alpha level of 0.05;

1. There will be no significant relationship between easiness estimates obtained from two contrasting score groups of the same standardisation sample
2. There will be no significant relationship between ability estimates obtained from two contrasting score groups of the same standardisation sample.
3. There will be no significant relationship among the item parameters obtained from the one parameter model, the two-parameter model and the classical item point biserial correlation co-efficient.
4. There will be no significant relationship between the classical raw scores and its logit exponential transformation.

The literature review gives some justification for applying the latent trait models in the development and standardisation of a physics achievement test for senior secondary physics student. However, questions regarding unidimensionality, the fit of the data to the model, construct validity of the data and the amount of error and ability necessary to achieve a given raw score, remain unanswered. This study provides answers to the above questions using a physics achievement test. The sample consisted of 2215 senior secondary school III physics students randomly composed from 62 senior secondary schools in Rivers State, employing simple random sampling, stratified random sampling and cluster sampling techniques.

The instrument for the study consisted of 60, five-option multiple- choice question on a physics achievement test. It included back ground items that asked for demographic data. Demographic data were gathered on the following subjects characteristics; age, sex, occupation of father and school location. The over-all reliability co-efficient (KR-20) of the physics achievement test was 0.89. the development of the PAT items proceeded in a number of identifiable stages; planning, preparing, pilot testing and evaluating. The PAT items were submitted to three experts for careful editing and critical review in order to establish the PAT face validity. To determine the PAT content validity, the items were checked with the table of specification to ensure that the test showed the desired proportion of emphasis. The content validity of PAT was also established with the assistance of experts, who were required to match each item with the instructional objectives. The experts also ascertained the correct answer for each item.

The PAT was administered to the subjects following a standardized procedure. The subjects were required to shade one of the five symbols by each item that most appealed to them. The items were scored right or wrong (one or zero) and the raw score was the total number of correct answers. Data analysis was performed using the PROX algorithm and regression methods on the Microsoft Excel visual BASIC computer programme. The chi-square goodness-of-fit test was used to investigate how well the latent trait model fits the PAT items. Finally, factor analysis based on the varimax option was used to establish the unidimentionality of the final version of the PAT items. Results of the study show that:

1. The factor analysis of the physics achievement test results in 14 Eigen values greater than 1.0
2. An analysis of model fit shows that two items failed to fit the model.
3. The overall reliability co-efficient (KR-20) of the physics achievement test was 0.89
4. There was no significant relationship, statistically, between easiness estimates obtained from two contrasting score groups of the same standardisation sample.
5. There was a significant relationship between ability estimates obtained from the contrasting score groups of the same standardisation sample
6. The inter-correlation among validity indices obtained from the one-parameter mode, the two- parameter model, the classical point biserial and the factor loading were high and statistically significant.
7. The correlation between ability estimates based on the classical and latent trait models was nearly perfect.

From the findings of the study, it was concluded that:

1. The physics achievement test measured one single trait for physics ability
2. The subjects generally responded in a consistent manner on the physics achievement test, as evidenced by the compatibility of the physics achievement test to the latent trait model.
3. The physics achievement test is valid and reliable
4. The latent trait model easiness estimates are not invariant across groups of examinees that differ in ability
5. The latent trait model analysis leads to invariant ability estimates while the classical test model analysis does not.

6. There is a considerable overlap among validity indices obtained from the one-parameter model, the two-parameter model and the classical point biserial correlation co-efficient.

Among the implications of the study, it was pointed out that the findings of the study might be addressed to those responsible for item banking. The study also has implications for latent trait investigations of item bias, for counselling, tailored measurement and criterion referenced testing. It was recommended that the latent trait analysis should be followed up by some qualitative analysis as some item misfits cannot be determined alone by statistical results. It was further recommended that statistical treatment of items that showed a misfit should be investigated by the use of the nominal response model. One factor which reduced the invariance of the item parameters and limits the generalisability of the findings of the study is that subjects with low scores were included in the validation process.

REFERENCES

Adesina, S. (1984). *Nigerian Educational Trends and issues.* Ile-Ife: University of Ife Press Ltd.

Agwagah, U.V.N. (1985). The Development and Prelimary Validation of an Achievement Test in Mathematics methods. *Unpublished Master's thesis, University of Nigeria, Nsukka.*

Allen, M.J.& Wendy, M . (1979). *Introduction to Measurement Theory.* California: Wadsworth, Inc.

Allen, N.L., Ansley, T.N. & Forsyth, R.A (1987). The Effect of Deleting content - related items on IRT Abilities Estimates. *Educational and Psychological measurement* 47 (4), 1141 - 1152

Ali, A. (1996). *Fundamentals of Research in Education.* AWKA: Meks Publishers (Nig.)

Anastasi, A (1976). *Psychological Testing (2edn)* Newyork:Macmillan Publishing Company

Anderson, E B . (1973). A Goodness of fit Test for the Rasch Model.*Psychometrika,* 38, 123 - 140.

Anderson, J., Kearney, G.E. & Everett, A.V (1968). An Evaluation of Rasch's Structural Model for tests of Intelligence. *British Journal of Mathematical and statistical Psychology,* 21, 231 - 238

Andrich, D. (1980). Using Latent Trait Measurement Models to analyse Attitudinal Data. In spearits, D.(ED). *The Improvement of Measurement in Education and Psychology: Contributions of Latent Trait Theories.* Australia: Australian Council for Educational Research.

Anene, G.U. & Ndubuisi, O.G. (1992) Test Development Process. In Nworgu, B.G. (Ed). Educational Measurement and Evaulation: *Theory and practice.* Nsukka: Hallman Publishers

Angoff, W.H. (1971). Scales, Norms, Equivalent Scores. In Thorndike, R.L. (Ed). *Educational measurement.* Washington, D.C. : American Council on Education.

Baker, F.B.(1977). Advances in Item Analysis. *Review of Educational Research,* 47, 151 - 178

Baker, F.B. & Martin, J.J. (1969). FORTRAN Test Analysis Package. *Journal of Educational and Psychological measurement,* 29, 159 - 164.

Bartholome, D.J. (1987). *Latent Variable Models and Factor Analysis.* Newyork: Charles Griffin and Company Ltd.

Bentler, B. M. (1980). The study of Cognitive Development Through Casual Modeling with Qualitative Data. In Kluwe, I & Spada, H.(Eds), *Development Models of Thinking.* Newyork: Academic Press.

Benson, J., Crocker. L., & Ware, W.A. (1978) Comparison of three types of item analysis in Test Development using classical and Latent Trait methods. *ERIC. ED* 178,556.

Best, J.W.(1981). *Research in Education (4th edu.).* Engle- wood clifts, Newjersy: Prentice Hall Inc.

Birenbaun, M. (1985): Comparing the effectiveness of several IRT based appropriateness measures in delecting unusual response patterns. *Educational and Psychological measurement,* 45(3-4), 523-539.

Birnbaun, A. (1968) Some latent trait models and their use in infering an examinee's ability. In Lord, F. M. & NOVICK, M.R. (Eds). *Statistical theories of mental test scores.* Reading: Addison - Wesley.

Bloom, B.S. (1976). *Human characteristics and school learning.* Newyork David Mckay Co.

Bloom, B.S., Englechart, M.D., Furst, B.J., Hill, W.W. & Krath-Wohl, D.R. (1980). *Taxonomy of Educational objectives the classification goals,* Newyork: David Mckay Company.

Bock, R.D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika,* 37 29 - 51.

Bock, R.D. & Lieberman, M (1970). Fitting a response model for dichotomously scored items. *Psychometrika,* 35, 179 - 197.

Bock, R -D. & Wood, R. (1971). Test theory. In P.H. Mussen & M.R. Rosenwing (Eds). *Annual Review of psychology,* 22, 620-645

Borg,, W.& Gall, M.D. (1983). *Educational Research: An Introduction* Newyork: David Meckay Co. Inc.

Browne, M.W. (1968). A comparision of Factor Analytic Techniques *Psycholometrika*, 33 (3), 267 - 298.

Bright, B.D. & Douglas, G.A(1977). Conditional Versus Unconditional procedures for sample - free analysis. *Educational and Psychological measurement,* 37, 573 - 586.

Bryce, T.G.K. (1981). Rasch Fitting. *British Educational Research Journal,* 7(2), 137 - 153.

Campbell, D.T. & Fiske, D.W. (1959). Convergent and Discriminant validation by the multi-trait, Multi - method matric. *Psychology Bulletin*, 56, 81 - 105

Chapple, M.(1976). *"O" & "A"* Level physics. London: The chaucer Press Ltd.

Choppin, B.H. (1980). The Use of Latent Trait Models in the Measurement of Cognitive Abilities and Skills. In spearitt, D.(Ed.) *The improvement of measurement in Education and psychology*: Contributions of latent trait theories. Austratia: Australian Council for Educational Research.

COX, D.R. (1970): *The Analysis of Binary data.* London: Chapman and Hall.

Cressie, N (1983). Characterizing the Manifest Probabilities of Latent Trait models *psychometrike,* 48(1), 129 - 141.

Cronbach, L.J. (1951). Co-efficient Alpha and the Internal Structure of tests. *Psychometrika,* 16, 297 - 334

Cronbach, L.J. (1971). Test Validation. In Thorndike, R.L. (Ed). *Educational Measurement* (2edn). Washington, D.C.: American Council on Education.

Cronbach, L.J. Rajaratanam, N. & Glesser, G. (1969). Theory of generalizability: A liberalization of reliability theory. *British Journal of statistical psychology*, 16 137 -163.

Dixton, W.J. & Massey, F.J. (1969) *Introduction to Statistical Analysis (3rd edn)*. Newyork: McGraw Hill Book Company.

Douglas, G.A.(19980). Conditional Inference in a Generic Rasch Model. In spearitt, D. (Ed). *The improvement in Education and Psychology: Contributions of latent trait theories.* Australia: Australian Council for Educational Research.

Douglas, G.A.(1990). Latent Trait measurement models. In Keeve, J.P.(Ed). *Educational Research, Methodology and measurement: An International Handbook.* Newyork: Pergamon press.

Douglas, F.M., Khavari, KA. & Farber, P.D. (1979). A comparison of classical and Latent Trait Item Analysis Procedures *Journal of Educational and Psychological measurement,* 36, 337 - 351.

Dowine, M.N. & Health, R.W. (1974) *Basic Statistical Methods (4th edn).* Newyork: Harper & Row Publishers.

Drasqow, F. & Lissak, I. (1983). Modified Parallel Analysis: A Procedure for examining the latent Dimensionality of Dichotomously scored item responses. *Journal of Applied Psychology,* 68, 363 - 373.

Doubois, P.H. (1970). A History of psychological Testing. In Allen J.M. & Wendy, M. (1979). *Introduction to Measurement Theory.* California: Wadworth, Inc.

Ebel, R.L. (1972). *Essentials of Educational Measurement (2nd edn)* Englewood cliffs: Prentice-Hall

Ezebor, S (1983). *Tests, Evaluation and Performance in Nigeria.* Ile -Ife: University of Ife Press Ltd.

Fafunwa, J.A.(1974) *History of Education in Nigeria.* London: George Allen Unwin.

Federal Ministry of Education, Science & Technology (1985). *Handbook on continous Assessment.* Ibadan: Heinemann (Nig) Ltd.

Ferguson, G.A. (1942). Item Selection By the Constant Process. *Psychometrika,* 7, 19 - 29.

Finney, D.J. (1944). The Application of Probit Analysis to the Results of mental tests. *Psychometrika,* 9, 31 - 39.

Fischer, G.H. (1973). The Linear Logistic Model as an instrument in Educational Research. *Acta Psychologia,* 37, 359 - 374.

Fischer, G.A.(1976). Some Probabilistic Models for Measuring change. In Degruijther, D.M. & Vander-Kamp, L.J. (Eds). *Advances in psychological and Educational Research.* Newyork: Wiley.

Fraser, B. J.(1980). Development and Validation of a Test of Enquiry Skills. *Journal of Research in Science Teaching,* 17(1), 7 - 16.

Gagne, N.L. (1988). *The Scientific Basis of the Art of Teaching.* Newyork: Teachers College Press.

Gallini, J.K. (1983) A Rasch Analysis of Raven Item Data. *The Journal of Experimental Education,* 52(1), 27 - 32.

Glass, G.V. (1966). *Note on Rank Biserial Correlation Co-efficient.* Journal of Educational and psychological measurement, 26, 622 - 631.

Goldman, S.H. & Raju, N.S.(1986). Recovery of one - and Two - parameter Logistic Item parameters. An Empirical Study. *Journal of Educational and Psychological measurement,* 46 , 11 - 21.

Green, S.B. Lissitz, R.W. & Mulark, S.A (1977). Limitation of Co-efficient Alpha as An Index of Test Unidimensionality. *Educational and Psychological Measurement,* 37, 827 - 838.

Grondlund, N.E. (1980). *Measurement and Evaluation in Teaching* (3rd Edn.) Newyork: Macmillan Publishing Company Inc.

Guilford, J.P. (1954) *Psychological Methods (2 edn.)* Newyork: McGraw-Hill.

Guiton, R.M. & Ironson, G.H.(1983). Latent Trait Theory for organisational Research. *Organisational Behaviour and Human performance,* 31, 54 - 87.

Gulliksen, H.(1950) *Theory of mental Tests.* Newyork.

Guttman, L (1945). The basis for analysis of test/retest reliability. *Pschometrika,* 10, 255 - 282.

Hambleton, R.K. & Cook, L.L.(1977) Latent Trait Models and their use in The Analysis of Education test data. *Journal of Educational Measurement,* 14, 75 - 96.

Hamblejon, R .K., Swaminathan, H., Cook, L.L., Eignor, D. R. & Guilford, J.A (1978). Development in Latent Trait Theory: Models, Technical issues and applications. *Review of Educational Research,* 48, 467 - 510.

Harrison, D.A. (1986). Robustness of IRT Parameter Estmation to violations of the unidimensionality assumption. *Journal of Educational Statistics,* 11 (2), 91 - 115.

Harwell, M.R.(1983). A compasnsion of two item selection procedures in criterion-referenced measurement. *Unplished Doctoral Dissertation,* University of Wisconsin - Madison.

Henryson, S.(1962). The relation between factor loadings and biserial correlation in item analysis. *Psychometrika,* 27, 419 - 424.

Henryson, S (1971). Gathering, analysing and using data on text items. In Thorndike, R.L(Ed). *Educational measurement.* Washingston, D.C: American Council on Education.

Herman, J.(1990) . Item writing Techniques. In Walberg, H.J. & Haertel, G.D. (Eds). *The International Encyclopedia of Educational Evaluation.* Newyork: Pergomon press.

International Centre for Educational Evaluation (1982). *Education Research*in Nigeria. Institute of Education: University of Ibadan.

Izard, J.F. & White, J.D.(1980). The use of Latent Trait Models in the Development and Analysis of classroom 'Tests. In spearitt, D (Ed). *The improvement of measurement in Education and Psychology: contributions of latent trait theories.* Austrlian Council for Educational Research.

Joe, I.A.(1992) Fundamental Statistics for Education and Behavioural Sciences. Ibadan: Krait Books Limited.

John, W.L., Annabel, J.M., Ryan, J.M. & Edens, K.M.(1995). A Rasch Analysis of factors derived from the Charles, F.K. School Climak profile. *Educational and Psychological measurement,* 55 (3), 456 - 567.

Keats, J.A. (1980). Comparing Latent trait with classical measurement models in the practice of Educational and psychological measurement. In Spearitt, D (Ed). *The improvement of in Measurement in Education and Psychology: Contribution Latent Trait Theories.* Australia: Austrilan Council for Educaional research.

Kelly, T.L. (1939) The selection of upper and lower groups for the validation of test items. *Journal of Educational Psychology,* 30 , 17 - 24.

Kerlinger, F.N. (1973). Foundation of Behavioural Research (2nd edn). Newyork: Holt, Rine-Hart and Wiston.

Kim, T.N. (1979), Applicability of classical Test score models to repeated perfromance on the same test. *Australlian Journal of Psychological Measurement,* 26 , 1 - 8.

Korashy, A.F.(1995). Applying the Rasch model to the selection of items for a mental ability Test. *Educational and Psychological, Measurement,* 55(5), 753 - 763.

Kuang, H.P.(1952) critical Evaluation of relative efficiency of three techniques in item Analysis. *Educational and Psychological measurement*, 39, 248 - 266.

Kuder, G.F. & Richardson, M.V. (1937). The theory of estimation of test reliability. *Psychometrika,* 3, 151 - 160.

Lawley, D.N. (1943). On problems connceted with item selection and test construction. In Baker, F.B. (1977), Advances in items Analysis. *Review of Educational Research,* 47, 151 - 178.

Lawshe, C.H.(1942). A monograph for estimating validity of test items. *Journal of applied psychology,* 26, 846 - 849.

Lazarshfeld, P.F. (1950). *The logical and mathematical Foundation of latent Structure Analysis.* Newyork, Houghton Mifflin.

Leabo, D.A. (1972). *Basic Statistics*, Georgetown, Ontario; Richard D. Irwin Inc.

Lemk, E. & Wiersmal, W.(1976) *Principles of Psychological measurement.* Chicago:: Rand McNally.

Lewis, D.G.(1974). *Assesment in Education.* London: University of London on Press Ltd.

Lindvall, C.M. & Nitko, A. (1975). *Measuring Pupil Achievements and Aptitudes.* Newyork: Harcourt Bruce Inc.

Livingston, S.A.(1988). Reliability of Test results. In Keeves, J.P. (Ed) *Educational Research, Methodology and measurement:* An International Hand bool. Newyork: Pergamon Press.

Lord, F.M.(1968). An analysis of the verbal scholastic Test using Birbaun is three-parameter logistic-model. *Educational Psychological measurement,* 28, 989 - 1020.

Lord, F.M.(1980). *Applications of Item response theory to practical testing problems.* Hillsdale, NJ: Lawrence Erlbaum Assoc, publishers.

Lord, F.M. & Novick, M.R.(1968). *Statistical theories of mental test scores* Reading, Mass: Addison-Welly.

Lumsden, J.(1961). The Construction of unidimensional tests. *Psychological Bulletins*, 58 122 - 131.

Lumsden, J.(1976). Test theory *Annual Review of Psychology*, 27, 251 - 280.

Martua, V.R.(1977). *Applying norm - Referenced and criterion - Referenced measurement in Education.* London: Allyn and Bacon, Inc.

Mckenzie, D.L. & Padilla, M.J. (1986). The Construction and validation of the test of Graphing in science (TOGS). *Journal of Research in Science Teaching*, 23(7), 571-579.

Mehrens, W.A. & Lehmann, I.I.(1978). *Measurement and Evaluation in Education and Psychology (2nd edn).* Newyork; Holt, Rinchart & Winton.

Morgan, G. (1980). The use of the Rasch Latent Trait Measurement Model in the Equating of Scholastic. Aptitude Tests. In spearitt, D (Ed). *The improvement of measurement in Educaion and Psychology Contribution, of latent Trait theories.* Australia: Australian Council for Educational Research.

Ndukwe, J.O. (1090). The Development and Preliminary Validation of an Integrated Science Achievement Tes (ISAT) *Unpublished Master's thesis,* University of Nigeria, Nsukka.

Neil, J.A. & Jackson, D.N.(1970). An Evaluation of item selection strategies in personality scale construction. *Educational and Psychological measurement*, 30, 647 - 661.

Nelkon M. & Parker, P. (1968). *Advance level physics.*, London: Heinemann.

Nkemakolam, O.C. (1992). Development and validation of an Instrument for measuring Test Anxiety among secondary school students. *Unpublished Doctoral Dissertation,* University of Nigerian, Nsukka.

Nworgu, B.G. (1985). The Development and Preliminary validation of a physics Achievement Test (PAT). *Unpublished Master's thesis, University of Nsukka.*

Nworgu, B.G. & Harbor-Peters, V.F.A. (1990). The development and validation of physics achievement test (PAT). *Journal of Science Instruction and Curriculum Studies*, 15, 33-39,

Obioma, G.O. (1985). The Development and validation of a Dianostic Mathematics Achievement Test for Nigeria Secondary School Students. *Unpublished Doctoral Dissertation, University of Nigeria*, Nsukka.

Ogochukwu,F.N.(1990). The Development and validation of an Instrument (Shyness Scale) for the identification of shyness. *Unpublished Master's thesis,* University of Nigeria, Nsukka.

Ogomaka, P.M.C.(1984) Development and Preliminary validation of students Evaluation of Teachers Effectiveness Scale. *Unpublished master's thesis, University* of Nigeria, Nsukka.

Ohuche, R.C. & Akeju, S.A. (1977). *Testing and Evaluation in Education.* Lagos: African Educaional Resources.

Okeke, P.N. & Anyakoha, M.W.(1987). *Senior Secondary Physics .* London: Macmillan Publishers Ltd.

Ponterotto, J.G. Burkard, A. & Rieger, B.P. (1995). Development and Initial validation of the Quick Discrimination index. *Educational and Psychological measurement,* 55(6), 1016 - 1031.

Popham, W.J.(1971). *Indices of adequacy for criterion - Referenced measurement: An Intrudction.* Englewood cliffs: Educational Technology publications.

Rao, C.R.(1956). Estimation and Tests of significance in factor analysis. Psycholometrika, 20, 92 - 111.

Rasch, G. (1966). An item Analysis which takes individual differences into account. *British Journal of Mathematical and statistical Psychomogy,* 19(1) 49 - 57.

Rastogi, K.G.(1983). *Educationa Psychology.* India: Rastogi Publications.

Reckase, M.D.(1979). Unifactor Latent trait models applied to mulifactor tests: Results and implications. *Journal of Educational statistics,* 4 (3), 207-230.

Reiser, M.(1983). An Item Response model for the estimation of Demographic effects *Journal of Educational Statistics,* 8 (3) 165 - 186.

Ross, K. (1966). An empirical study of a logistic mental test model. *Psychomsetrika,* 31, 325 - 340.

Samejima, F(1969). Estimation of latent ability using a reponse pattern of graded scores. *Psychometric monograph, No. 17, Psychometrika,* 34, part 2, 1 - 97.

Samejima, F.(1972). A general model for free-response data. *Psychometric Monograph,* No. 18.

Sax, G.(1989) Principles of Educational and psychological *measurement and evaluation.* Callifornia: Wadsworth Publishing Company.

Scates, T.(1947). In coffman, W.(1969). *Achievement Tests.* A Project of the American Educational Research Association.

Schofield, H.(1978). *Assessment and Testing: An Introduction.* London: George Aken.

Schontz, F.C. (1965). *Research Methods in Personality.* Newyork; Appleton -century Grofts.

Smith, R.M.(1996). A comparison of the rasch separate calibration and between fit methods of Detecting item bias. *Educational and psychological measurement,* 55(3), 403 - 417.

Spada, H. & May,R.(1980). The linear logistic test model and its Application in Education Research. In spearitt, D.(Ed), *The improvement of measurement in Education and Psychology; contributions of latent trait Theories.* Australia: Australian Council for Educational Research.

Soriyan, M.A.(1972). Measurement of the goodness of fit of Rasch probabilistic model for item analysis to objective achievement test of the West African Certificate Examination. *Dissertation Abstracts International 32(83) 4433A*

Stanley, J.C.(1971). Reliability. In Thorndike, R.L.(Ed). *Educational Measurement (2nd edn).* Washington, D.C.: American Council on Education.

Susan, Ed.(1983). Construct Validity: Construct Representation versus Nomothetic Span. *Psychological Bulletin,* 93, 179 - 199.

Thorndike, R.L. (1990). Reliability. In Keeves, J.P.(Ed). *Educational Research, Methodology, and Measurement. An International Handbook:* Newyork: Peramon Press.

Thorndike, R.M.(1990). Correlational rocedures. In Keeves, J.P.(Ed). *Educational Research, Methodology and Measurement:* An International *Handbook.* Newyor: Pergamon Press.

Tinsley, H.E. & Dawis, R.V.(1975). An investigation of the Rasch simple logistic model, sample free item and test calibration. *Education and Psychological measurement,* 35, 325 - 339.

Tobin, K. & Caple, W.(1982). Development and Validation of a group test of integrated science processes. *Journal of Research in Science Teaching,* 19(2), 133 - 141.

Tucker, L.R. (1948). A method of scaling ability test items, taking item unreliability into account. *American psychologist,* 3, 509 - 510.

Urry, V.M.(1974). Approximation to item paramekers of mental test models and their uses. *Educational and psychological measurement* 34, 253 - 269.

Vanderlinden, W.J.(1981). A latent trait look at the pretest - Post test validation criterion - referenced test items. *Review of Educational Research,* 51, 371 - 402.

Wainer, M. & Bright B. (1980). Robust estimation of ability in the Rasch model. *Psychometrika,* 45(3), 373 - 39.

Wainer, H., Morgan, A., & Gustafson, J.(1980). A review of estimation procedures for the Rasch model with an eye toward longist test. *Journal of Educational Statistics,* 5, 35 - 64

Weiss, D.T. & Davisson, M.L.(1981). Test theory, and methods. In Rosenzweig, M.H. Porter, L.W.(Eds). *Annual Review of Psychology,* 32, 629 - 651.

Wesman, A.G.(1971). Writing the test items. In Thorndike, R.L. (Ed). *Educational measurement (2nd edn).* Washingston, D.C.; American Council on Education.

Wilcox, R(1979). Achievement Tests and latent structure models. *British Journal of Mathematical and Statistical Psychology.* 32, 61 - 67.

Wood, R.(1990). Item analysis. In Keeves, J.P.(Ed). *Educational Research Methodology and measurement.* Newyork: Pergamaon Press.

Wright, B.D.(1967). Sample - free test calibration and person measurement *Educational and Psychological measurement,* 35, 325 - 339.

Wright, B.D.(1977). Solving measurment problems with the Rasch model. *Journal of Educational Measurement,* 14, 97 - 116.

Wright, B.D. & Panchapakesan, N.(1969). A procedure for sample - free item analysis. *Educational and Psychological Measurement,* 29, 23 - 48.

Youngman, M.B.(1979). A comparison of item - total pointbiserial correlation, Rasch's and Alpha Beater item analysis procedures. *Educational Studies,* 5(3), 265 - 273.

APPENDIX I

UNIVERSITY OF NIGERIA, NSUKKA
FACULTY OF EDUCATION
SUB-SCIENCE DEPARTMENT OF EDUCATION

PHYSICS ACHIEVEMENT TEST (FINAL FORM)

Number: .... Sex: ....

Age: .... School:..

Local Govt. Area: ....

Occupation of Father:..

N/B: Information will be used for research purpose **only**.

Instruction:

1. Time Allowed: 1 hr.
2. Attempt all questions
3. Answer each question by choosing one correct letter and shading it.

Example

When a force moves through a certain distance, it is said that:

A. Work have been done
B. Energy have been loss
C. Power have been loss
D. Energy have been gained
E. Energy have been converted from one form into another.

The correct answer is A, and A is therefore shaded.

Now do the following:

1. A quantity which possesses attribute of magnitude and direction is called

A. Mass
B. Force
C. Scolar (Knowledge)
D Power
E. Vector

2. A canon ball is projected so as to attain a maximum range. Find the maximum height attained if the initial velocity is U.

A. $\frac{2u \sin \theta \cos \theta}{g} \frac{U \sin \theta}{g}$

B. $\frac{U^2 \sin 2\theta}{g}$

C. $\frac{2U \sin \theta}{g}$ (Application)

D. $U^2/2g$

E. $\frac{2u \sin \theta \cos \theta}{g}$

3. Which of the following statement(s) is/are true of inertia?

I Inertia is the rate of change in velocity

II Inertia is the product of mass x velocity

III Inertia is the numerical measure of mass

IV Inertia is that tendency of a body to remain at rest or to move with uniform velocity.

A. I only

B. II and III only

C. III and IV only (Knowledge)

D. III only

E. IV only

4. A bullet of mass 0.020kg and velocity 500mls pierces a stationary wooden block of mass 1.25kg. Find the velocity of the block immediately after the collision.

A. 7.87 mls
B. 8.13 mls
C. 10.00 mls (Application)
D. 62.50 mls
E. 625.0 mls

5. Two blocks of the same dimensions, one, steel and the other wooden, are dropped simultaneously from the same height. If they fall freely, neglecting air resistance, the

A. Two blocks hit the ground simultaneously because their dimensions are the same.
B. Two blocks hit the ground simultaneously because they have the same acceleration.
C. Wooden block, being lighter than the steel block reaches the ground first. (Comprehension)
D. Steel block reaches the ground first because it is denser than wooden block.
E. Steel block takes half as much time as the wooden block to reach the ground because it is more massive than the wooden block.

6. A machine has a velocity ratio of 5 and is 80% efficient, what effort would be needed to lift a load of 2000N with the aid of this machine?

A. 80N
B. 160N
C. 400N
D. 500N (Application)
E. 1600N

7. The below table was constructed after rolling a ball down an incline plane.

Time(s)	0	1	2	3	4	5	6
Distance(m)	0	1	4	9	16	25	36

The formular expressing the relationship between D and T is

A. D is directly proportional to T^2
B. D is inversely proportional to T^2
C. D is inversely proportional to T (Synthesis)
D. D is directly proportional to 2T
E. D is directly proportional to ½T

8. Which of the following is true of density measurement?

A. Architects and engineers make use of density measurement in the design of bridges.
B. Argon at high pressure is used in gas filled with electric lamps
C. Relative density has units such as kg/m^3 (Comprehension)
D. The density of a gas is independent of temperature and hence the pressure of a gas.
E. Balloons used in weather soundings in the upper atmosphere or in cosmic-ray investigation are filled with a heavy gas.

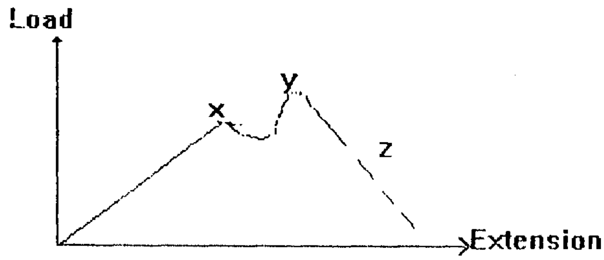
9. The following statements are true of friction except

A. It opposes motion.
B. It can be reduced by rollers (Comprehension)
C. Kinetic friction is smaller than stratic friction.
D. Without friction we would be unable to work.
E. Friction is independent of the nature of surface.

10. The magnitude of stress per unit strain of solid a material is the

A. Modulus of elasticity.
B. Bulk modulus
C. Shearing force (Knowledge)
D. Youngs modulus
E. Tensile force.

11.



In the diagram above, the point y is called

A. Elastic limit
B. Yield point (Comprehension)
C. Breaking point
D. Inelastic limit
E. Shearing point

12. When water is spilled on a clean glass surface, it wets the glass. This is because

A. Adhesion of water to glass is stronger than the cohesion of water.
B. Cohesion of water to glass is stronger than the adhesion of water.
C. There is a force of attraction between molecules of water and glass.
D. The water surface is not sufficiently concave (Knowledge)
E. Water behave as an elastic skin.

13. A 5-kg mass does not fall five times as fast as a 1kg mass of the same size and shape because

A. the earth gravitational force is the same for both
B. the ratio of force to mass is the same for both
C. air friction has more effect on the 5kg mass
D. when two masses have the same kinetic energy the lighter one moves faster
E. all masses fall at the same rate in a vacuum. (Knowledge)

14. A force of 10N another of 12N may be applied so that their resultant is

A. 0.00N
B. 2.00N
C. 15.62N (Comprehension)
D. 22.00N
E. 120.00N

15. If two objects x and y have the same momentum, y can have more kinetic energy only if it

A. Weighs more than X
B. Weighs the same as X
C. is moving faster than X (Analysis)
D. is moving more slowly than X
E. occupies a position above that of X

16. A student in attempting to determine the relative density of methylated spirit used the following materials: relative density bottle, methylated spirit, glycerine, water, beam balance, blotting paper and the spring balance. What was wrong with this experimental set up?

A. Using the spring balance to measure weight of bottle instead of the beam balance.

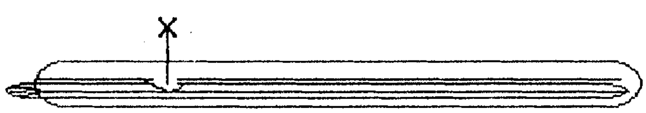
B. Using the beam balance to measure weight of bottle instead of the spring balance.

C. Comparing the mass of methylated spirit with an equal volume of water instead of glycerine.

D. Using blotting paper to wipe excess water on the bottle instead of warming the bottle.

E. First weighing the bottle empty instead of weighing if filled with methylated spirit. (Evaluation)

17.



Clinical thermometer

The function of x above is to

A. Measure the pressure of the mercury

B. Callibrate the thermometer

C. Remove dissolved impurities from the mercury (Knowledge)

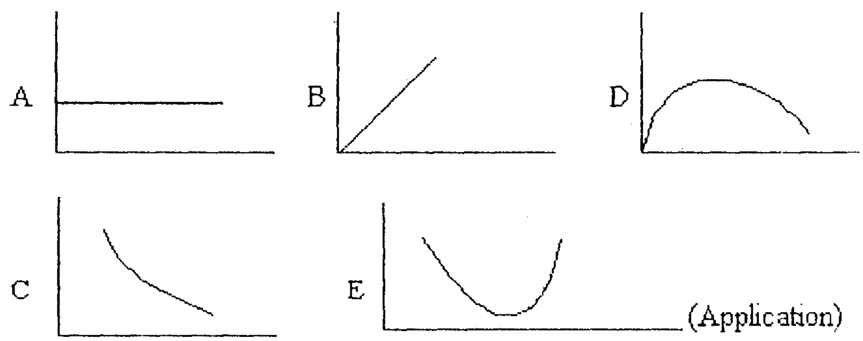
D. Regulate the flow of mercury along the bulb.

E. Assist in the expansion of mercury.

18. Water contract between the temperature

A. $-0^{O}C$ to $0^{O}C$
B. $-0^{O}C$ to $4^{O}C$
C. $-4^{O}C$ to $0^{O}C$ (Comprehension)
D. $0^{O}C$ to $4^{O}C$
E. $4^{O}C$ to $10^{O}C$

19. Which of the following diagrams correctly represents the variation of the pressure P of a fixed mass of gas with volume V



(Application)

20 If the pressure of a constant volume of gas is 30cm of mercury at $20^{O}C$, find its pressure at $30^{O}C$

A. 28.00 CmHg
B. 63.00 CmHg
C. 77.38 CmHg (Application)
D. 87.20 CmHg
E. 195.00 CmHg

21. Which of the following statements is not true of the expansion of solid.

A. Creaking noises are heard due to iron expanding as it gets hot.
B. A metal structure such as a bridge expands when heated.
(Comprehension)
C. To prevent expansion, gaps are left between section of rail on a railway track.
D. If hot water is poured into a thick glass bottle, it is liable to track as a result of uneven expansion of the glass.
E. We can remove the tight stopper of a glass bottle without cracking either the bottle or the stopper due to expansion of the glass bottle.

22. Which of these is true of saturation vapour pressure.

A. Saturation vapour pressure obeys gas laws.
B. Saturation vapour does not easily liquefy
(Comprehension)
C. A saturation vapour pressure is a pressure which is in disequilibrium with its liquid.
D. At saturation vapour pressure, rate of evaporation is unrelated to the rate of condensation.
E. A saturation vapour pressure is a pressure which is in equilibrium with its liquid.

23. Heat from the sun reaches us mainly by which of the following processes.

A. Conduction
B. Transmission (Knowledge)
C. Convection
D. Radiation and convection
E. Radiation.

24. In the Thermo flask, the silvered walls prevent heat loss by:

A. Conduction only
B. Conduction and conversion
C. Radiation only (Knowledge)
D. Conduction and radiation only
E. Convection and radiation only.

25. A collection of rays of light is known as

A. Shadows
B. Eclipses (Knowledge)
C. Ream
D. Rectilinear propagation
E. Converging light

26. Which of the following is not a characteristic of the image in a plane mirror.

A. It is the same size as the object
B. It is far behind the mirror as the object is in front
C. It is upright (Knowledge)
D. It is laterally inverted
E. It is upright

27. When an object O is placed in front of two mirrors M_1 and M_2 inclined at an angle of 90^o. The number of images observed is

A. 2
B. 3 (Comprehension)
C. 4
D. 6
E. 8

28. The moon travels around the earth once every month, but we are unable to see the eclipse every month. The reason is that:

A. the earth is spherical
B. the moon is non-luminous (Comprehension)
C. the earth rotates on its axis
D. the plane of the moon orbit is not the same as that of the earth around the sun
E. the earth is not a perfect sphere.

29. The convex mirror is used as a driving mirror because

I It forms real images of real objects

II It has a relatively large field of view

which of the following is/are correct?

A. I only
B. I and III only
C. I and II only (Comprehension)
D. II only
E. II and III only

30. An object is placed 4cm in front of a converging lens of focal length 10cm produces a magnified image 7cm high. Find the position of the image and nature of the image.

A. 2.25cm and virtual
B. 2.25cm and real
C. 4.20cm and real (Application)
D. 6.70cm and virtual
E. 6.70cm and real.

31. In an experiment for the formation of a pure spectrum, a student used the following materials a narrow slit as a source of light, a diverging lens, a 60" prism for dispersion, a converging lens, and a screen. What as wrong with this experimental set up.

A. Using a narrow slit instead of a wide slit (Evaluation)
B. Using a diverging lens instead of a converging lens.
C. Using a 60^{o} prism instead of a 45^{o} prism
D. Using a converging lens instead of a diverging lens
E. Using a screen instead of an aperture.

32. A sound wave emitted from the bottom of a ship travels at 1500mls vertically downwards through water to an ocean bed 100m below, is reflected upwards. What is the time interval between the instant the sound is emitted and the instant the echo is received?

A. 0.67 seconds
B. 1.33 seconds
C. 1.50 seconds (Application)
D. 3.00 seconds
E. 15.00 seconds

33. In a transverse wave propagation, the distance between two adjacent crest is

A. twice the wavelength
B. one quarter the wavelength
C. three quarter of the wavelength (Comprehension)
D. the wavelength
E. half the wavelength

34. In a resonance tube experiment, a tuning fork of frequency 256Hz gave resonance when the water level was 32.5cm below the open end of the tube. If the next position of resonance was 100cm, what is the speed of sound in air?

A. 256.0mls
B. 278.7mls (Application)
C. 317.2mls
D. 333.2mls
E. 345.6mls

35. An instrument has three identical strings which are subjected to the following tensions, 5N, 10N, 20N. It was observed that the string which has tension 20N have the highest frequency. The formula expressing the relationship between frequency and the tension in the string is

A. frequency is proportional to the tension
B. frequency is proportional to the square of the tension (Analysis)
C. frequency is proportional to the square root of the tension
D. frequency is inversely proportional to the square of the tension
E. frequency is inversely proportional to the square root of the tension.

36. A glass cylinder opened at both ends produces a fundamental note. If the velocity of sound in air is V and L the length of the pipe, which of the following expresses the relationship between the frequency of the note, the velocity of sound and the length of the pipe (neglect end-correction)

A. frequency = 2V/L
B. frequency = V/2L
C. frequency = V/L
D. frequency = V/4L
E. frequency = V/5L

37. A shadow is produced when light passes an obstacle. The shadow has the same shape as the obstacle. This is because light travels

A. through vacuum
B. in straight lines
C. a short distance
D. a long distance
E. through air

38. Total internal reflection will not occur when light travel from

A. air into glass
B. glass into air
C. glass into water (Knowledge)
D. water into air
E. air into vacuum

39. The colour of light which is least deviated is

A. violet
B. Green
C. Red (Knowledge)
D. Indigo
E. Blue

40. A long sighted person can see objects at a distance but cannot see close objects distinctly. This suggests that his

A. near point is shorter than the normal near point which is 25cm
B. eyeball is too long
C. eyelens is not sufficiently divergent (Comprehension)
D. near point is further away than the normal near point which is 25cm
E. eyelens is too convergent

41. A $2^{U}F$ capacitor is charged by connecting it across a battery of emf 100 volts. What energy is available from the discharge of the capacitor.

A. $1 \times 10^{-12}J$
B. $1 \times 10^{-10}J$
C. $1 \times 10^{-6}J$ (Application)
D. $1 \times 10^{-4}J$
E. $1 \times 10^{-2}J$

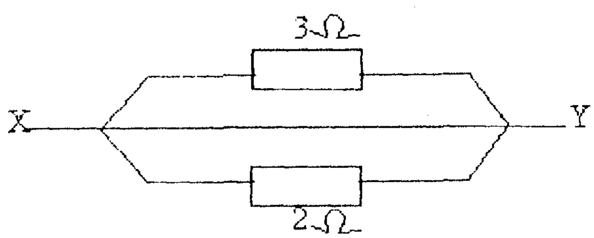
42. Which of the following is not true of the electrical resistance of a wire?

A. The resistance is proportional to the length of the wire.
B. The resistance increases with increasing area of the wire

(Comprehension)

C. The resistance depends on the nature of the materials.
D. The resistance of a given wire increases with increased temperature
E. The resistivity of a material is the resistance between opposite faces of a unit cube made from the material.

43. The diagram below shows two resistors connected in parallel between the points X and Y. The single resistor which would have the same resistance between X and Y is



(Application)

A. 0.67Ω
B. 0.83Ω
C. 1.00Ω
D. 1.2Ω
E. 5.00Ω

44. A galvanometer is converted to voltmeter by

A. Connecting a shunt in parallel with the instruments
B. Connecting a shunt in series with the instrument (Knowledge)
C. Connecting a multiplier in series with the instrument
D. Connecting a multiplier in parallel with the instrument
E. Connecting a shunt and multiplier in parallel with the instrument

45. The heating effect of an electrical current is used in fuses to safeguard circuit against excessive currents. Fuse is used for this purpose because

A. fuse depends on the expansion of a resistance wire
B. fuse has high melting point
C. fuse has high boiling point (Comprehension)
D. fuse has low melting point
E. fuse is independent of temperature

46. The angle between the direction of magnetic North and direction of geographical North at a particular point is called.

A. angle of declination
B. angle of dip (Knowledge)
C. angle of meridian
D. neutral angle
E. angle of inclination

47. The neutral point around a bar magnet is

A. The point at which the magnetic field due to a bar magnet is maximum
B. The point at which the magnetic field due to a bar magnet is zero
C. The point at which the magnetic field due to a bar magnet points in the North-South direction (Comprehension)
D. The point at which the resultant magnetic flux density is zero
E. The point at which the magnetic field due to a bar magnet is in the same direction with the earth magnetic field.

48. The relation between the direction of the magnetic field and the electric current through a conductor is given by

A. Fleming left-hand rule
B. Fleming right-hand rule (Knowledge)
C. Faraday's law of electro magnetic induction
D. Maxwell's cork-screw rule
E. Lenz's law

49. Which of the following is not true of the Bohr's atomic model?

A. An atom consists of nucleus and a number of electrons
B. Most of the masses of an atom is concentrated in the nucleus
C. The nucleus is positively charged (Knowledge)
D. The nucleus is central and extremely large
E. The electrons are held in the outer parts of the atom by the force of attraction.

50. Electrons of an atom have the following characteristics except:

A. They carry a negative charge
B. They travels in straight line (Knowledge)
C. They are extremely light
D. They have momentum
E. They have no energy

51. The atomic weight of an element is not a whole number. This is due to the fact that

A. an atom of an element contains three particles namely electrons, protons and neutrons (Comprehension)
B. atoms have a structure which varies periodically with increasing weight
C. The atomic weight of an element is the number of protons in its nucleus added to the number of neutrons
D. The protons and neutrons are both called necleons
E. Atoms of any element exists in isotopic form.

52. A nucleus whose atomic number and neutron number are known is called.

A. Nucleus
B. Mass number (Knowledge)
C. Nuclide
D. Isotopes
E. Atomic weight

53. An alpha-particle decay is described by the equation $^{A}_{Z}X \rightarrow {}^{A-4}_{Z-2}Y + {}^{4}_{2}He$ is called

A. Mass number (Knowledge)
B. Parent nuclide
C. Parent nucleus
D. Daughter nucleus
E. Parent nucleus

54. A radioactive element has 4.0×10^8 atoms initially. Calculate the remaining number of atom in 32 minutes if the decay constant is 0.043/s

A. 1.0×10^8
B. 2.0×18^8
C. 4.0×10^8 (Application)
D. 4.8×10^8
E. 6.0×10^8

55. The half life of an element is 16 years. Calculate the time it will take 32g of the element of decay to 2g.

A. 4 years
B. 8 years (Application)
C. 16 years
D. 32 years
E. 64 years

56. Which of the following statements concerning the behaviour of diode valves is/are true

I. Valve conducts electron when the anode potential (Va) is negative
II. Valve conducts electron if the anode potential (Va) is positive
III. The magnitude of the saturation current increases with the temperature of the cathode.

A. I only (Comprehrension)
B. III only
C. I and II only
D. I and III only
E. II and III only

57. The following statements are true of cathode rays except:

A. They cause glass and other materials to flourescence
B. They are negatively charged particles flowing along the tube
C. They are deflected by magnetic field
D. They are electromagnetic radiation
E. They have momentum. (Comprehension)

58. In a given medium, the expression relating the force of attraction or repulsion, F between two bodies with electrons e_1 and e_2 and separated by a distance, Y is given by

A. $F = \frac{1}{4}\pi_{\varepsilon o}\frac{ei}{\gamma}$

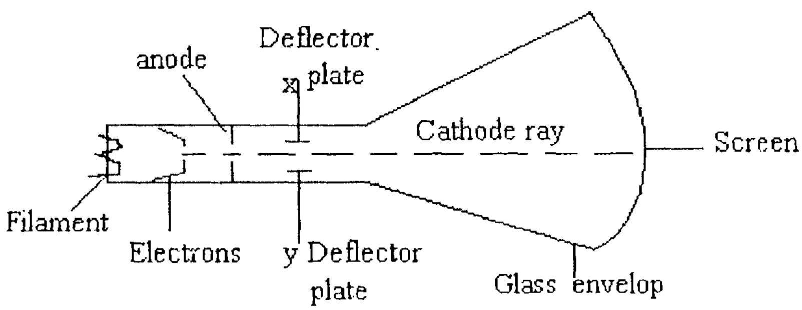
B. $F = \frac{1}{4}\pi_{\varepsilon o}\frac{ei\,e2}{\gamma}$

C. $F = \frac{1}{4}\pi_{\varepsilon o}\frac{e2}{\gamma^2}$ (Analysis)

D. $F = \frac{1}{4}\pi_{\varepsilon o}\frac{e1\,e2}{\gamma^2}$

E. $F = \frac{1}{4}\pi_{\varepsilon o}\frac{e1}{\gamma^2}$

59



The figure above shows a simple cathode ray tube with only one pair of deflector plates, x and y. What would be the effect of connecting the plates to a battery?

A. The rays will be deflected electro statically (Evaluation)

B. The rays will be deflected magnetically

C. The temperature of the filament will be raised

D. The plate x will be negatively charged while y positively charged

E. The anode voltage will be increased while the plates remain charged.

60. If A is the mass number of an atom and z its atomic number, the α-particle decay of the atom can be described by the following equation

A. ${}^{A}_{Z}x \rightarrow {}^{A-4}_{Z-2}Y + {}^{4}_{2}He$

B. ${}^{Z}_{A}X \rightarrow {}^{4-A}_{Z-2} + {}^{0}_{1}ne$

C. ${}^{A}_{Z}X \rightarrow {}^{A-4}_{Z-2} + {}^{4}_{2}He$ (Synthesis)

D. ${}^{Z}_{A}X \rightarrow {}^{A-4}_{Z-2} + {}^{4}_{2}He$

E. ${}^{A}_{Z}X \rightarrow {}^{A-4}_{Z-2} + {}^{H}_{2}He$

APPENDIX II

UNIVERSITY OF NIGERIA, NSUKKA
FACULTY OF EDUCATION
SUB-SCIENCE DEPARTMENT

<u>PHYSICS ACHIEVEMENT TEST (FIRST VERSION)</u>

Represents correct option.

1.

What is the correct reading of the venier caliper shown above

A. 4.30cm
B. 4.32cm (Application)
C. 5.25cm
D. 5.00cm
E. 4.31cm

2. A quantity which possesses attribute of magnitude and direction is called

A. Mass
B. Force
C. Scalar
D. Power (Knowledge)
E. Vector

3. A car moves from rest with an acceleration of $0.2mls^2$. Its velocity as it moved a distance of 50m is

A. 4.47mls
B. 10.00mls (Application)
C. 50.20mls
D. 50.00mls
E. 1.00mls

4. The velocity with which a body hits the ground after falling 50m from rest is

A. 50mls
B. 500mls (Application)
C. 30mls
D. 31mls
E. 50.00mls

5. A canon ball is projected so as to attain a maximum range. Find the maximum height attained if the initial velicity is u

A. $\frac{2\ u^2 \sin\theta\ \text{Cos}\,\theta}{g} \frac{\text{Usin}\,\theta}{g}$

B. $\frac{U^2 \sin 2\theta}{g}$

C. $\frac{2\ u\sin\theta}{g}$ (Application)

D. $U^2/2g$

E. $\frac{2\ u^2 \sin\theta\ \text{Cos}\,\theta}{g}$

6. Which of the following statement(s) is/are true of Inertia

I Inertia is the rate of change in velocity

II Inertia is the product: mass x velocity

III Inertia is the numerical measure of mass

Iv Inertia is that tendency of a body to remain at rest or to move with uniform velocity

A. I only

B. II and Iv only

C. III and Iv only (Evaluation)

D. III only

E. IV only

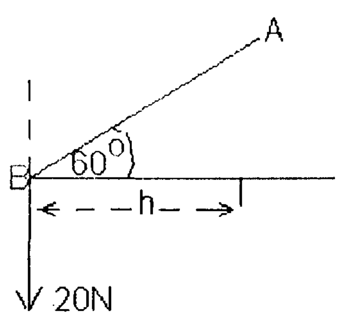
7. An object of mass 5kg moves round a circle of radius 4m with a speed of 10mls. What is its angular velocity?

A. 0.5 $rads^{-1}$
B. 0.10 $rads^{-1}$
C. 10.00 rad^{-1} (Application)
D. 2.50 $rads^{-1}$
E. 40.00 $rads^{-1}$

8 . A boy on a bicycleis going downhill without pedaling. He notice that he moves faster if he puts his head below the handle bars, this is because

A. He puts more effort on the front wheel
B. He reduces the resistance of his body to the air
C. He helps the wheels to go faster (Appiication)
D. He increases the pressure on the pedals
E. He increases the resistance of his body to the air.

9.



Find the moment of the 20N force about the points A and B.

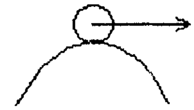
A. 60 Nm (Analysis)
B. 20 Nm
C. 30 Nm
D. 52 Nm
E. 1.5 Nm

10. A bullet of mass 0.020kg and velocity 500mls pierces a stationary wooden block of mass 1.25kg. Find the velocity of the block immediately after the collision.

A. 7.87 mls
B. 8.13 mls
C. 10.00 mls (Application)
D. 62.50 mls
E. 625 mls

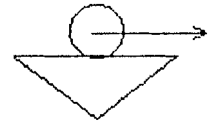
11. Which of the following figures represent a body in neutral equilibrium?

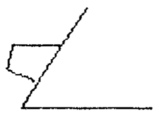
A. 

B. 

(Comprehension)

C 

D. 

E. 

12. Two blocks of the same dimensions one steel and the other wooden, are dropped simultaneously from the same height. If they fall freely, neglecting air resistance, the

A. Two blocks hit the ground simultaneously because their dimension are the same.
B. Two blocks hit the ground simultaneously because they have the same acceleration
C. Wooden block, being lighter than the steel block reaches the ground first.
D. Steel block reaches the ground first because it is denser than the wooden block.
E. Steel block takes half as much time as the wooden block to reach the ground because it is more massive than the wooden block. (Comprehension)

13. The pitch of the screw of a screw jack is 0.5 cm when used to raise a load, the handle turns through a circle of radius 40cm. Determine the mechanical advantage of the jack it the efficiency is 25% ($\prod = 3.14$)

A. 502.4
B. 3.14
C. 314 (Application)
D. 250
E. 125.6

14. In which of the following machine is the effort applied between load and fulcrum?

A. Scissors
B. Pilers
C. Wheel barrow (Application)
D. Mut cracker
E. Human arms

15. A machine has a velocity ratio of 5 and is 80% efficient, what effort would be needed to lift a load of 200N with the aid of this machine?

A. 80N
B. 160N
C. 400N (Application)
D. 500N
E. 1600N

16. The below table was constructed after rolling a ball down an incline plane

Time(s)	0	1	2	3	4	5	6
Distance (m)	0	1	4	9	16	25	36

The formular expressing the relationship between D and T. is

A. D is directly proportional. to T^2
B. D is inversely proportional to T^2
C. D is inversely proportional to T
D. D is directly proportional to the constant T
E. D is directly proportional to ½T (Sythesis)

17. A brass bolt in weighs 0.69N. If the relative density of brass is 8.6. Calculate the apparent weight of the bolt in water (Density of water = $1000kg/m^3$).

A. 5.98N
B. 598N
C. 0.081N (Application)
D. 0.08N
E. 860N

18. Which of the following is true of density measurements?

A. Architects and engineers make use of density measurement in the design of bridges
B. Argon at high pressure is used in gas filled with electric lamps.
C. Relative density has units such as kg/m^3.
D. The density of a gas is independent of temperature and hence the pressure of a gas
E. Balloons used in weather soundings in the upper a atmosphere or in cosmoc-ray, investigations are filled with a heavy gas

(comprehension)

19. The following statements are true of friction except

A. It opposes motion
B. It can be reduced by rollers (Comprehension)
C. Kinetic friction is smaller than static friction.
D. Without friction we would be unable to work
E. Friction is independent of the nature of surface.

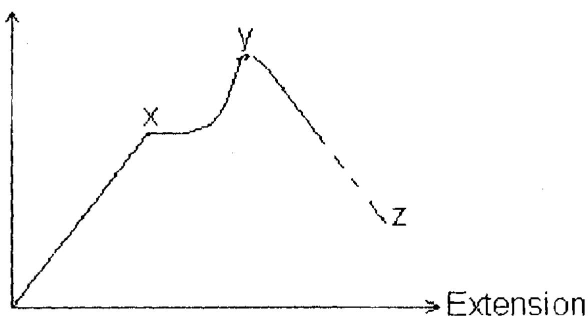
20. A body of mass 40kg is given an acceleration of $10mls^2$ on a horizontal ground for which the co-efficient of friction is 0.5. Determine the magnitude of the force required to accelerate the body (Take g = $10mls^2$).

A. 400N
B. 50N
C. 600N (Application)
D. 100N
E. 40N

21. The magnitude of stress per unit strain is the

A. Modulus of elasticity
B. Bulk modulus
C. Shearing force (Knowledge)
D. Young's modulus
E. Tensite force

22.



In the diagram above, the point y is called

A. elastic limit
B. Yield point
C. Breaking point (Comprehension)
D. Inelastic limit
E. Shearing point.

23. A spring of natural length 3m is extended by 0.01m by a force of 4N. What will be its length when the applied force is 12N?

A. 4.03m
B. 3.04m
C. 3.03m (Application)
D. 3.10m
E. 0.03m

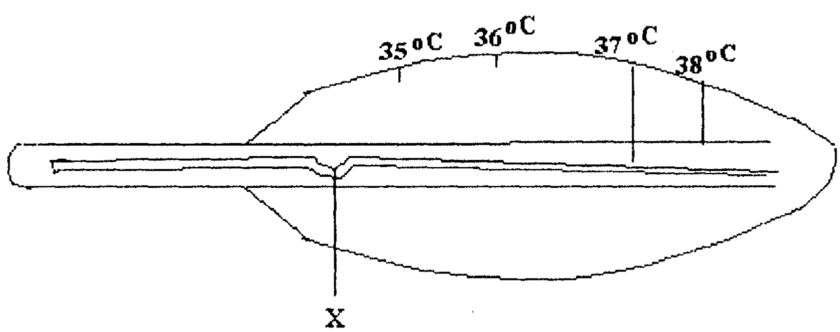
24. When water is spilled on a clean glass surface it wets the glass. This is because

A. Adhesion of water to glass is stronger than the cohesion of water.
B. Cohesion of water to glass is stronger than the adhesion of water.
C. There is a force of attraction between molecules of water and glass.
D. The water surface is not sufficiently concare.
E. Water behave as an elastic skin. (Knowledge)

25. Mercury is preferred to alcohol as a thermometric liquid because

A. Mercury boils at 78^{o}C and hence can be used to measure high temperature
B. Vaporises easily (Knowledge)
C. Freezes at - 39^{o}C and hence cannot be used for measuring very low temperature
D. Freezes at - 115^{o}C and hence can be used for measuring very low temperatures
E. High expansivity.

26.



Clinical thermometer

The function of X above is

A. To measure the pressure of the mercury
B. To calibrate the thermometer (Knowledge)
C. To remove dissolved impurities from the mercury
D. To regulate the flow of mercury along the bulb
E. To assist in the expansion of mercury.

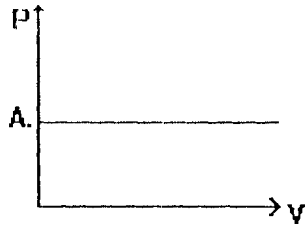
27. A piece of lead falls 3m from rest, coming to rest again on the ground. Calculate the rise in its temperature, the specific heat capacity of lead being 0.130J/gk or 130J/kgk

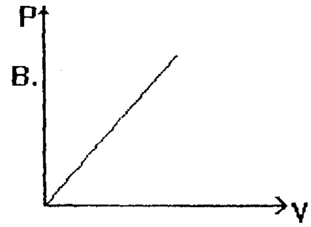
A. $390^{o}C$
B. $0.39^{o}C$ (Application)
C. $0.23^{o}C$
D. $23^{o}C$
E. $30^{o}C$

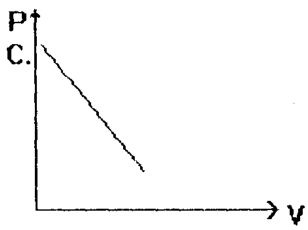
28. Water contract between the temperature

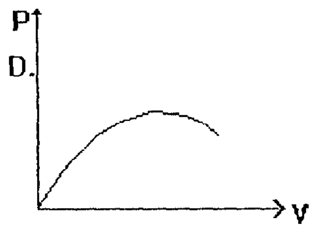
A. $-0^{o}C$tp $0^{o}C$
B. $0^{o}C$ to $4^{o}C$
C. $4^{o}C$ to $10^{o}C$
D. $-4^{o}C$ to $0^{o}C$ (Comprehension)
E. $-0^{o}C$ to $4^{o}C$

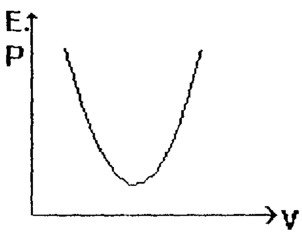
29. Which of the following diagrams correctly represents the variation of the pressure P of a filled mass gas with volume, V?











(Application)

30. If the pressure of a constant volume of gas is 70cm of mercury at $20^{o}C$, find its pressure at $30^{o}C$.

A. 175 cmHg
B. 63.50 cmHg
C. 77.20 cmHg
D. 28.00 cmHg
E. 87.20 cmHg

(Application)

31. Which of the following statements is not true of the expansion of solid.

A. Creaking noises are heard due to iron expanding as it gets hot.
B. A metal structure such as a bridge expands when heated.
C. To prevent expansion, gaps are left between section of rail on a railway track.
D. If hot water is poured into a thick glass bottle, it is liable to track as a result of uneven expansion of the glass.
E. We can remove the tight stopper of a glass bottle without cracking either the bottle or the stopper due to expansion of the glass bottle. (Comprehension)

32. The coefficient of linear expansivity of a metal is 0.000019 per K. What will the area of $400mm^2$ become if its temperature is raised by 10^oK.

A. $400.152mm^2$
B. $400\ mm^2$ (Application)
C. $400.760mm^2$
D. $4000\ mm^2$
E. $190\ mm^2$

33. In an experimental determination of the apparent cubic expansivity:

Mass of liquid expelled from bottle = $(M_2 - M_3)$

Mass of liquid left in bottle = $(M_3 - M_1)$

Temperature rise = $(\theta_2 - \theta_1)$

The apparent cubic expansivity, y is calculated from the relationship

A. $y = \dfrac{(\theta_2 - \theta_1)(M_2 - M_1)}{(M_3 - M_1)}$

B. $y = \dfrac{M_2 - M_3}{(M_2 - M_3)(\theta_2 - \theta_1)}$ (Synthesis)

C. $y = \dfrac{\theta_2 - \theta_1}{(M_2 - M_3)(M_3 - M_1)}$

D. $y = \dfrac{(\theta_2 - \theta_1)(M_3 - M_1)}{(M_2 - M_1)}$

E, $y = \dfrac{(\theta_2 - \theta_1)(M_2 - M_1)}{(M_3 - M_1)}$

34. An iron rod has length 50m when the temperature is 20^{o}C. By how much will it expand when the temperature rises to 30^{o}C?

A. 50.00 6m
B. 0.00 6m
C. 50.60 m (Application)
D. 0.60 m
E. 10.60 m

35. The specific latent heat of fusion of ice is 336000J/kg and the specific heat capacity of water is 4200J/kg. Calculate the heat required to convert 0.250kg of ice at its melting point to water at $20^{o}C$

A. 21000J
B. 1050J
C. 10100J
D. 126000J
E. 84000J (Application)

36. A sample of air at $20^{o}C$ and a pressure of 75mmHg is found to have a dew point of $14^{o}C$. If the values of the saturation vapour pressure at $20^{o}C$ and $14^{o}C$ are 17./5 and 12.00 mmHg respectivity. Calculate the relative humidity.

A. 69.0%
B. 146.0%
C. 70.0% (Application)
D. 143.0%
E. 14.3%

37. Which of the following is true of saturation vapour pressure?

A. Saturation vapour pressure obeys gas laws.
B. Saturation vapour does not easily liquefy.
C. A saturation vapour pressure is a pressure which is in disequilibrium with its liquid.
D. At saturation vapour pressure, rate of evaporation is unrelated to the rate of condensation.
E. A saturation vapour pressure is a pressure which is in equilibrium with its liquid. (Comprehension)

38. Heat from the sun reaches us mainly by which of the following processes?

A. Conduction
B. Transmission (Knowledge)
C. Convection
D. Radiation and Convection
E. Radiation.

39. In the thermos flask, the silvered walls prevent heat loss by

A. Conduction only
B. Conduction and Convection
C. Radiation only
D. Conduction and radiation
E. Convection and radiation only (Knowledge)

40. Which of the following phenomena clearly demonstrates evaporation

A. A noisy phenomenon
B. Is dependent of external pressure (Comprehension)
C. Causes cooling
D. Atmospheric pressure does not assist evaporation
E. Evaporation is independent on area of exposed liquid.

41. A collection of rays of light is known as

A. Shadows
B. Eclipses
C. Beam (Knowledge)
D. Rectilinear propagation
E. Converging light

42. Materials like window glass which transmit a large proportion of light are said to be

A. Rays
B. Opague
C. Non-luminous (Knowledge)
D. Transparent
E. Translucent

43. Which of the following is not a characteristic of the image in a plane mirror.

A. It is the same size as the object
B. It is far behind the mirror as the object is in front (Knowledge)
C. It is upright
D. It is laterally inverted
E. It is upright.

44. When an object O is placed in front of two mirrors M_1 and M_2 inclined at an angle of 90". The number of images observed is

A. 3
B. 4
C. 2
D. 8
E. 6 (Comprehension)

45. Total internal reflection will not occur when light travel from

A. Air into glass
B. Glass into air
C. Glass into water (Comprehension)
D. Water into air
E. Air into vacuum

46. The colour of light which is least deviated is

A. Violet
B. Green
C. Red (Comprehension)
D. Indigo
E. Blue

47. The reflection of a narrow beam of light incidence normally on a plane mirror falls on a metre rule placed parallel to the mirror and at a distance of 75cm from it. If the mirror is rotated through 5^{O}C by what distance is the reflected beam displaced along the metre rule?

A. 6.56cm
B. 6.54cm
C. 14.72cm
D. 13.22cm (Application)
E. 13.86cm

48. A shadow is produced when light passes an obstacle. The shadow has the same shape as the obstacle. This suggests that light travels:

A. Through vacuum
B. In straight lines (Comprehension)
C. A short distance
D. A long distance
E. Through air.

49. The Moon travels around the earth once every month, but we are unable to see the eclipse every month. The reason is that

A. The earth is spherical
B. The moon is non-luminous
C. The earth rotates on its axis
D. The plane of the moon orbit is not the same as that of the earth around the sun.
E. The earth is not a perfect sphere. (Comprehension)

50. The convex mirror is used as driving mirror because

I It forms real images of real objects

II It collects light from a wide angle

III It has a relatively large field of view

A. I only
B. I and II only
C. I and III only (Comprehension)
D. II only
E. II and III only

51. An object 0.4m high is placed in front of and at right angles to the axis of a concave mirror which has a radius of curvature 0.15m. If a real image 0.12m high is obtained, determine the distance of the object from the mirror

A. 10cm
B. 5cm
C. 12cm (Application)
D. 15cm
E. 40cm

52. An object is placed 4cm is front of a converging lens of focal length 10cm produces a magnified image of 7cm high. Determine the position of the image.

A. 4.2 cm and real
B. 6.7 cm and virtual
C. 6.7 cm and real (Application)
D. 2.25 cm and virtual
E. 2.25 cm and real

53. In an experiment for the formation of a pure spectrum, a student used the following materials; a narrow slit as a source of light, a diverging lens, a 60^{o} prism for dispersion, a converging lens, and a screen. What was wrong with this experimental set up.

A. Using a narrow slit instead of a wide slit.
B. Using a diverging lens instead of a converging lens
C. Using a 60^{o} prism instead of a 45^{o} prism
D. Using a converging lens instead of a diverging lens
E. Using a screen instead of an aperture. (Evaluation)

54. A longsighted person can see objects at a distance but cannot see close objects distinctly. What is wrong with this person?

A. His near point is shorter than the normal near point which is 25cm
B. The eyeball is too long
C. The eyelens is not sufficiently divergent
D. His near point is further away than the normal near point which is 25cm
E. The eyelens is too convergent. (Evaluation)

55. A wave travels a distance of 80m in 4 seconds. The distance between successive crests of the wave is 50cm. Find the frequency of the wave.

A. 20H2
B. 30H2
C. 50H2
D. 40H2
E. 80H2 (Application)

56. A sound wave emitted from the bottom of a ship travels at 1500mls vertically downwards through water to an ocean bed 1000m below, is reflected upwards. What is the time interval between the instant the sound is emitted and the instant the echo is received?

A. 0.67s
B. 1.5s
C. 1.33s
D. 3s
E. 15s (Application)

57. In a transverse wave propagation, the distance between two adjacent crest is

A. Twice the wavelength
B. One-quarter the wavelength
C. Three-quarter of the wavelength (Comprehension)
D. The wavelength
E. Half the wavelength

58. In a resonance tube experiment, a tuning fork of frequency 26H2 gave resonance when the water level was 32.5cm below the open end of the tube. If the next position of resonance was 100 cm, what is the speed of sound in air?

A. 278.7mls
B. 317.28mls (Application)
C. 333.92mls
D. 256mls
E. 345.6mls

59. An instrument has three identical strings which are subjected to the following tensions: 5N, 10N, 20N. It was observed that the string which has tension 20N have the highest frequency. The formular expressing the relationship between frequency and the tension in the string is

A. Frequency is proportional to the tension
B. Frequency is proportional to the square of the tension
C. Frequency is proportional to the square root of the tension
D. Frequency is inversely proportional to the square of the tension
E. Frequency is inversely proportional to the square root of the tension. (Synthesis)

60. A glass cylinder opened at both ends, produces a fundamental note. If the velocity of sound in air is V and L the length of the pipe, which of the following expresses the relationship between the frequency of the note, the velocity of sound and the length of the pipe (neglect end-correction).

A. Frequency = 2V/L
B. Frequency = V/2L
C. Frequency = V/L
D. Frequency = V/4L (Synthesis)
E. Frequency = V/5L

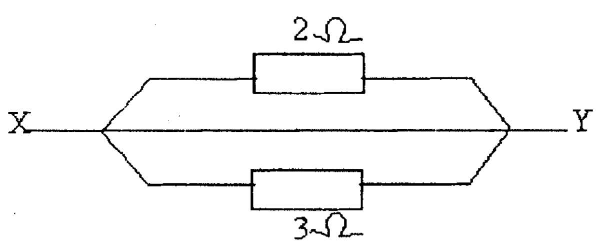
61. A 2UF capacitor is charged by connecting it across a battery of emf 100 volts. What energy is available from the discharge of the capacitor.

A. 1×10^{-10}J
B. 1×10^{-12}J
C. 1×10^{-2}J
D. 1×10^{-6}J
E. 1×10^{-4}J (Application)

62. Which of the following is not true of the electrical resistance of a wire?

A. The resistance is proportional to the length of the wire.
B. The resistance increases with increasing area of the wire (Comprehension)
C. The resistance depends on the nature of the materials.
D. The resistance of a given wire increases with increased temperature
E. The resistivity of a material is the resistance between opposite faces of a unit cube made from the material.

63. The diagram below shows two resistors connected in parallel between the points X and Y. The single resistor which would have the same resistance between X and Y is



(Application)

A. 0.67Ω
B. 0.83Ω
C. 1.00Ω
D. 1.2Ω
E. 5.00Ω

64. A galvanometer is converted to voltmeter by

A. Connecting a shunt in parallel with the instruments
B. Connecting a shunt in series with the instrument
C. Connecting a multiplier in series with the instrument
D. Connecting a multiplier in parallel with the instrument
E. Connecting a shunt and multiplier in parallel with the instrument

(Knowledge)

65. The heating effect of an electrical current is used in fuses to safeguard circuit against excessive currents. Fuse is used for this purpose because

A. fuse depends on the expansion of a resistance wire
B. fuse has high melting point
C. fuse has high boiling point
D. fuse has low melting point (Comprehension)
E. fuse is independent of temperature

66. The angle between the direction of magnetic North and direction of geographical North at a particular point is called.

A. angle of declination
B. angle of dip
C. angle of meridian (Knowledge)
D. neutral angle
E. angle of inclination

67. The neutral point around a bar magnet is

A. The point at which the magnetic field due to a bar magnet is maximum
B. The point at which the magnetic field due to a bar magnet is zero
C. The point at which the magnetic field due to a bar magnet points in the North-South direction
D. The point at which the resultant magnetic flux density is zero (Comprehension)
E. The point at which the magnetic field due to a bar magnet is in the same direction with the earth magnetic field.

68. The relation between the direction of the magnetic field and the electric current through a conductor is given by

A. Flaming left-hand rule
B. Flaming right-hand rule
C. Faraday's law of electro magnetic induction
D. Maxwell's cork-screw rule (Knowledge)
E. Lenz's law

69. The induced emf in a circuits is directly proportional to the rate of the magnetic flux or field lines linking the circuit. This is a statement of

A. Lenz's law
B. Flexing right-hand rule
C. Flexing left-hand rule (Comprehension)
D. Faraday's law of electrolysis
E. Faraday's law of electromagnetic induction.

70. A circuit consists of a capacitor of $2^{U}F$ and a resistor of 1000Ω. An alternating emf of 12V(rms) and frequency 50HZ is applied. Find the current flowing in the circuit.

A. 1.690 A
B. 6.4×10^{-3}A
C. 1.59×10^{-3}A (Application)
D. 10.22×10^{-3}A
E. 6.04×10^{-2}A

71. The input emf of a transformer is 240V ac. The output emf is 960V a.c. If there are 720 turns in the secondary cost. Determine the number of turns in the primary cost.

A. 180 turns
B. 200 turns
C. 2080 turns (Application)
D. 320 turns
E. 302 turns

72. In a given medium, the expression relating the force of attraction or reputation, F between two bodies with charges, Q and q and separated by a distance, r is given by

A. F = 1/4Πgo
B. F = ¼Πgo
C. F = ¼Πgo (Analysis)
D. F = ¼Πgo
E. F = ¼Πgo

73. Which of the following is not true of the Bohr's atomic model?

A. An atom consists of nucleus and a number of electrons.
B. Most the masses of an atom is concentrated in the nucleus. (Knowledge)
C. The nucleus is positively charged.
D. The nucleus is central and extremely large.
E. The electrons are held in the outer parts of the atom by the force of attraction.

74. Which of the following statements is not true of the electron of an atom of an element.

A. They travel in straight lines
B. As heavy as hydrogen atom (Comprehension)
C. They have momentum
D. They produce flourescence in certain materials
E. They carry a negative charge.

75. The atomic weight of an element is not a whole number. This is due to the fact that

A. an atom of an element contains three particles namely electrons, protons and neutrons
B. atoms have a structure which varies periodically with increasing weight (Comprehension)
C. The atomic weight of an element is the number of protons in its nucleus added to the number of neutrons.
D. The protons and neutrons are both called nucleons
E. Atoms of an element exists in isotopic form.

76. A nucleus whose atomic number and neutron number are known is called

A. Nucleus
B. Mass number
C. Nuclide
D. Isotopes
E. Atomic weight (Knowledge)

77. An alpha – particle decay is described by the equation
${}^{A}_{Z}X \rightarrow {}^{A-4}_{Z-2}Y + {}^{4}_{2}Hc$ is known as

A. Mass number
B. Parent nuclide
C. Parent nucleus (Comprehension)
D. Daughter nucleus
E. Parent nucleous

78. A radioactive material has a half-life of 10 hours. What fraction of the original radioacitive nucel will remain after 30 hours

A. 1/10
B. 1/3
C. 1/8
D. 1/16
E. ½ (Application)

79. Two atoms of deuterium, 2H fused together to form helium 4H. If the masses of deuterium and helium are respectively 2.0147, and 4.0038mg. Calculate the energy released, given that the velocity of light is 3.0×10^8mls.

A. 5.42×10^{14}J
B. 1.8×10^{12}J
C. 23.4×10^{11}J
D. 23.4×10^{10}J
E. 23.4×10^{14}J (Application)

80 The number of atoms of a radioactive substance is initially 4.0×10^8. If the number reduces tgo 10×10^8 in 32 minutes, what is the half-life of the substance?

A. 4 mins
B. 10 mins
C. 64 mins
D. 8 mins
E. 12 mins (Application)

81. The half life of an element is 16 years. Calculate the time it will take 32g of the element to decay to 2g.

A. 8 years
B. 32 years (Application)
C. 64 years
D. 16 years
E. 4 years

82. Which of the following electro magnetic waves or radiation is produced when a high speed electron is stopped by a metallic target.

A. Gamma radiation
B. X-rays
C. Ultra-violet radiation (Comprehension)
D. Visible rays
E. Intra-red radiation

83. Which of the following statements concerning the behaviour of diode valves is/are true?

I Valve conducts electron when the anode potential (V_a) is negative
II Valve conducts electron if the anode potential (V_a) is positive
III The magnitude of the saturation current increases with the temperature of the cathode

A. I only
B. III only
C. I and II only
D. I and III only
E. II and III only (Comprehension)

84. The following statements are true of cathode rays excepts.

A. They cause glass and other materials to flourescence
B. They are negatively charged particles flowing along the tube
C. They are reflected by magnetic fields
D. They are electro-magnetic radiation
E. They have momentum (Comprehension)

85. Diodes act as rectifier because
A. Electrons flow freely from heated cathode to anode
B. Diodes do not obey Ohm's law (Comprehension)
C. Diodes obey Ohm's law
D. Electrons flow easily from anode to cathode
E. At saturation value, the current remains constant.

86. In the cathode ray oscilloscope, the cylindrical plates acts as

A. A source of electrons in the instrument
B. Screen on which spots are formed (Comprehension)
C. A mechanism to move the spot about the screen
D. The focus of the rays
E. A source of cathode ray

87. The work function of a metal is 1.9ev and it is illuminated by violet light of wave length 4.5×10^{-7} cm. Determine the maximum energy of liberated photo electron.

A. 1.9×10^{-19}J

B. 7.44×10^{-19}J

C. 1.36×10^{-19}J

D. 4.4×10^{-19}J

E. 3.04×10^{-20}J (Application)

APPENDIX III

A Chi-square goodness-of-fit statistic to test if there was a significant systematic bias in the choice of correct position of options for the 60 (final form) PAT items

	Frequency observed fo	Frequency expected, fe	Fo-fe	(fo-fe)2	(fo-fe)2/fe
A	11.00	12.00	-1.00	1.00	0.083
B	9.00	12.00	-3.00	9.00	0.750
C	13.00	12.00	1.00	1.00	0.083
D	17.00	12.00	5.00	25.00	2.083
E	10.00	12.00	-2.00	4.00	0.333
Total	60.00	60.00	0.00	40.00	3.332

The theoretical, expected frequency (fe) for each option

$$= \frac{11.0 + 9.00 + 13.00 + 17.00 + 10.00}{5} = 12.00$$

The null hypothesis (Ho) is that there is no significant bias in the choice of correct position of options. The criteria for reaching a decision at 0.05 level of significance are:

1. Accept Ho if $X_C^2 < 9.488$, 0.05% level of significance.
2. Reject Ho if $X_C^2 > 9.488$

The degree of freedom (df) = n-1 $\Rightarrow$ 5-1 = 4

Where n = 5, Number of options

Our $X^2_C = 3.33$

Therefore the null hypothesis is accepted.

Conclusion: The difference in the choice of correct position of options was not significantly biased from the statistical point of view.

APPENDIX IV
LETTER TO VALIDATORS

UNIVERSITY OF NIGERIA, NSUKKA
FACULTY OF EDUCATION
SUB-SCIENCE DEPARTMENT

Date:.....................

Dear Sir,

VALIDATION OF A PHYSICS ACHIEVEMENT TEST

Please, find attached, a 87-item multiple choice (five alternatives) Physics Achievement Test (PAT) and the cognitive domains being tested. The test items were written from the different content areas of physics using the Senior Secondary Physics curriculum.

Kindly establish the content validity, the correctness of answers (ticked option,) domain appropriateness and the overall clarity of the test items.

We solicit your comments on item clarity, objective clarity and the overall appropriateness of the items to allow for modifications, please.

Thank you.

Yours Sincerely

Signed
H.L. Nkpone

APPENDIX V

VALIDATION OF THE PHYSICS ACHIEVEMENT TEST (FIRST VERSION)

COMMENTS BY VALIDATORS

S/NO	ITEM NO.	FIRST VALIDATOR	SECOND VALIDATOR	THIRD VALIDATOR	REACTIONS
1	1	List item under application domain	Re-draw diagram to reflect equal spacing	No comment	Item listed under application. Diagram drawn to reflect equal spacing
2	3	No comment	Arrange options in either ascending or descending order	Include 20 mls as option	Numerical values of options arranged in ascending order. Included 20 mls as an option.
3	4	No comment	No comment	Supply the value of the accelerating due to gravity g = 9.8mls2	Value of g supplied (Take g = 9.8mls2
4	6	No comment	List item under analysis	No comment	Item listed under analysis
5	9	Application	Delete "and B"	Delete "and B"	"and B" deleted
6	11	No comment	Delete item	No comment	Item deleted
7	12	List under Knowledge	No comment	No comment	Item listed under knowledge
8	13	No comment	No comment	Take 12.56 as possible option	Included 12.56 as option
9	21	No comment	No comment	State if solid, liquid	Solid material
10	25	Structurally defective question, delete	No comment	Your keyed option is notan advantage as expected. Delete item	Item deleted
11.	27	No comment	No comment	Energy lost is gained as heat by both lead and ground, so the problem relies on a dangerous assumption. Delete	Item deleted
12	28	List item under Knowledge	No comment	No comment	Item listed under Knowledge
13.	34	No comment	No comment	State the coefficient of linear expansivity of Iron	Supplied the value for the coefficient of linear expansivit8y of Iron, that is 0.000011/°C
14.	37	List item under Knowledge	No comment	No comment	Item listed Knowledge
15.	40	Structurally defective Delete	No comment	Recast or delete	Item deleted
16.	48	No comment	Check the distracters.	State the question t hus: ... a shadow when an opaque object is in the path of light.	Item re-stated thus: a shadow is produced when an opaque object is in the path of light.
17	50	No comment	Change II to read: produces an erect image	No comment	Change alternative II to read "produces an erect image"

18	51	No comment	No comment	Re-write the item thus: "An object ... a concave mirror which has a radius of curvature	Item recast thus: "An object 0.04m high is placed in front of a concave mirror which has a radius of curvature 0.15m …….
19	52	No comment	No comment	Add "and nature of image to the last sentence in the stem"	"and nature of image added to the last sentence. In the stem.
20	61	No comment	No comment	The correct formular for energy - ½ C.V2 J and not ½VC2	Effected the correction in finding the energy of the discharged capacitor.
21	68	No comment	Insert "3through a conductor" after electric current	Sentence to read electric current through a conductor"	Item re-written as follows: "the electric current through a conductor is given by….
22.	69	No comment	Item should read, the magnitude...	No comment	Inse3rted magnitude "after the" to read the magnitude….
23.	70	No comment	No comment	State if capacitor is in series with resistor	Included that capacitor is in series with resistor
24.	72	No comment	Let the charges be q1 and q2 instead of Q and q	No comment	Stated that the charges be q1 and q2 respectively
25.	74	Recast	Revisit	No comment	Item recast thus electrons of an atom have the following characteristic except ….
26.	79	No comment	No comment	Delete, Question is on binding energy which is not in the syllabus	Item deleted
27.	80	No comment	No comment	Recast the item thus: A radicactive element has 4.0 X10^3 atom initially calculate the remaining atoms in 32 minutes if the decay constant is 0.043/S	Item re-written
28	87	No comment	Put item under modern physics	Supply the value of constants. Take h =6.62 x 10-34JS 1ev = 1.60x10-19C = 3 x 10 mls	Value of constant h, ev and velicity of light, C suplied as follows: h=6.62 x 10^{-34} J/S lev =1.60 x 10^{-19}C C = 3 x 10^8 mis

APPENDIX VI

PROX ITEM ANALYSIS PROCEDURES

Preliminary Procedures

Part I (Item)

1. The value, X, for each item is calculated thus:

$$X = \ln \frac{N - S}{S}$$

where N = Number of subjects
S = Total score on an item
ln = Natural logarithm

2. The mean of X is calculated thus:

$$\bar{X} = \frac{\sum x}{n}$$

where n = number of items

The square of x for each item = X^2, and the sum of $X^2 = \sum x^2$

3. The variance, U for the test is given as:

$$U = \frac{\sum x^2 - [n\, x(\bar{x})^2]}{n - 1}$$

Part II (Ability estimate)

1. For each subject total raw score, we calculate the value, b thus:

$$b = \ln \frac{\gamma}{n - \gamma}$$

where γ = subject total raw score
n = total number of items

2. The mean of b is calculated thus:

$\bar{b} = \frac{b}{N}$, and the square of b = b^2

where N = total number of subjects

3. The variance, V for the subjects score is given as:

$$V = \frac{\sum b^2 = [N \times (\bar{b})^2]}{N - 1}$$

Part III

1. Person's ability expansion factor (A) due to test width, range of item difficulties is given as:

$$A = \frac{1 + \frac{U}{2.89}}{1 - \frac{UV}{8.35}}$$

2. Item difficulty expansion factor (B) due to sample spread is given as:

$$B = \frac{1 + \frac{V}{2.89}}{1 - \frac{UV}{8.35}}$$

Part IV (Item validation)

(i) Difficulty index, for an item is given as:
$(x = \bar{x})$ multiplied by B.

(ii) The standard error, reliability for each item, is given as:

$$\text{Std. error} = \left[\frac{N}{S(N - S)}\right]^{1/2} \times B$$

(iii) Person measures ability estimate for each raw score is given as
$\theta j = b \times A$

(iv) The standard error associated with ability estimate, is given as:

$$\text{Standard error} = \left[\frac{n}{r(n-r)}\right]^{\frac{1}{2}} \times A$$

APPENDIX VII

MAXIMUM LIKELIHOOD ESTIMATION PROCEDURE (Unconditional)

In accordance with the latent trait test model:

$$P_g(\theta) = \frac{\text{Exp}^{\,Da_g(\theta - b_g)}}{1 + \text{exp}^{\,Da_g(\theta - b_g)}} \quad (g - -1,\ 2 - -n \text{ item}) \quad \text{------------------------(1)}$$

a_g = discrimination power for item g

b_g = difficulty index for item g

θ = subject ability

b = scaling factor = 1.7

$p_g(\theta)$ = Probability of answering item g, correctly

In additiion, let

n = the number of subjects

nr = the number of subjects with r, correct answers to an item

K = the number of items composed of the test.

The Likelihood function, L (Bock and Lieberman 1970) is:

$$L = \frac{n!}{\pi n!} \quad \frac{\pi^{2^k}}{K=1} \quad P_g(\theta)^{nr} \text{--------------------------------(2)}$$

π = Product function.

In order to obtain the maximum likelihood estimates of the parameter, the likelihood function is integrated with respect to each parameter of the validating sample. That is, $\alpha \log \gamma b_g$ = O, $\delta \log Li/\delta\theta_K$ = O, g = 1 = k -(3)

(for the one-parameter model), and

$$\frac{\delta \log L_2}{\delta ag} = O, \quad \frac{\gamma \log L_2}{\gamma ok} = O \qquad \text{---------------(4)}$$

(for the two-parameter model)

The likelihood functions, L_1 and L_2 are obtained by substituting the expression given by equation (1) for $Pg(\theta)$ in equation (2).

The simultaneous equations to be solved are:

1. K + n- 2 equations for the one-parameter model
2. 2k + n - 2 equations for the two-parameter model.

The simultaneous equations are solved by an interactive technique, the Newton-Raphson method (finding the second derivative of the log likelihood function).

The second derivative of the likelihood function in (equation 4) is:

$$\frac{\partial^2 \log L2}{\partial\theta_g \times \partial a_g} = 0, \quad \frac{\partial^2 \log L_2}{\partial b_g \times \partial b_g} = 0, \quad \frac{\partial^2 Log\, L_2}{\partial^2\ \theta k \times \partial 2\,\theta k}$$

These expressions are evaluated numerically to obtain the estimates of discrimination power (a_g), and difficulty index (b_g) for an item (Bock and Lieberman 1970).

APPENDIX VIII
Distractor Analysis

The distractor analysis was carried out to ascertain the number of students who choose the various options. For this purpose, the scores were ranked in order, from the highest to the lowest and the upper and lower 27 percent ability groups selected. The easiness (difficulty) and discrimination (distractor index) indices for each option were analysed, using the formular U+L/2N and U-L/N respectively.

Where U = Number in the upper 27 percent group who choose the option

L = Number in the lower 27 percent group who choose the option

N = Number of subjects in either the upper or lower 27 percent (598) group.

The results of the distractor analysis are presented below:

Alternative

Item		A	B	C	D	E	Omit
1	Easiness	0.03	0.05	0.05	0.13	0.74	0.00
	Discrimination	-0.06	-0.07	-0.06	-0.10	0.29	-0.01
2	Easiness	0.14	0.10	0.10	0.53	0.12	0.01
	Discrimination	-0.03	-0.13	-0.11	0.32	-0.12	-0.03
3	Easiness	0.11	0.10	0.60	0.08	0.10	0.00
	Discrimination	-0.19	-0.11	0.22	-0.00	-0.02	-0.00
4	Easiness	0.70	0.07	0.08	0.07	0.06	0.01
	Discrimination	0.33	-0.10	-0.07	-0.07	-0.06	-0.02
5	Easiness	0 05	0.66	0.09	0.12	0.08	0.01
	Discrimination	-0.04	0.37	-0.12	-0.11	-0.10	-0.01
6	Easiness	0.07	0.06	0.18	0.69	0.08	0.00
	Discrimination	-0.08	-0.06	-0.13	0.33	-0.08	-0.10

7	Easiness	$\boxed{0.46}$	0.35	0.09	0.07	0.03	0.00
	Discrimination	$\boxed{0.15}$	+0.01	-0.06	-0.07	-0.02	-0.01
8	Easiness	$\boxed{0.78}$	0.04	0.08	0.05	0.05	0.01
	Discrimination	$\boxed{0.22}$	-0.02	-0.09	-0.04	-0.05	-0.02
9	Easiness	0.08	0.07	0.06	0.08	$\boxed{0.70}$	0.10
	Discrimination	-0.03	-0.08	-0.06	-0.12	$\boxed{0.33}$	-0.02
10	Easiness	0.11	0.06	0.12	$\boxed{0.67}$	0.09	0.01
	Discrimination	-0.03	-0.06	-0.20	$\boxed{0.47}$	-0.16	-0.02
11	Easiness	0.05	$\boxed{0.81}$	0.05	0.05	0.05	0.00
	Discrimination	-0.05	$\boxed{0.18}$	-0.04	-0.07	-0.03	0.00
12	Easiness	$\boxed{0.68}$	0.06	0.08	0.08	0.07	0.03
	Discrimination	$\boxed{0.32}$	-0.05	-0.11	-0.07	-0.07	-0.02
13	Easiness	$\boxed{0.66}$	0.17	0.07	0.12	0.05	0.01
	Discrimination	$\boxed{0.24}$	-0.12	-0.08	-0.20	+0.01	-0.01
14	Easiness	0.09	0.05	0.07	$\boxed{0.69}$	0.09	0.01
	Discrimination	-0.14	-0.07	-0.07	$\boxed{0.34}$	-0.05	-0.01
15	Easiness	0.06	0.06	$\boxed{0.72}$	0.06	0.10	0.00
	Discrimination	-0.05	-0.01	$\boxed{0.19}$	-0.08	-0.07	-0.01
16	Easiness	$\boxed{0.69}$	0.04	0.06	0.08	0.12	0.01
	Discrimination	$\boxed{0.32}$	-0.03	-0.05	-0.12	-0.11	-0.01
17	Easiness	0.08	0.02	0.08	$\boxed{0.78}$	0.04	0.01
	Discrimination	-0.05	-0.02	-0.05	$\boxed{0.17}$	-0.04	-0.01
18	Easiness	0.10	0.04	0.05	$\boxed{0.77}$	0.05	0.00
	Discrimination	-0.05	-0.02	-0.07	$\boxed{0.17}$	-0.04	0.00
19	Easiness	0.09	0.05	$\boxed{0.75}$	0.05	0.05	0.01
	Discrimination	-0.09	-0.05	$\boxed{0.25}$	-0.06	-0.03	-0.02
20	Easiness	0.07	0.05	$\boxed{0.64}$	0.11	0.12	0.01
	Discrimination	-0.04	-0.03	$\boxed{0.12}$	-0.04	-0.01	-0.02
21	Easiness	0.05	0.05	$\boxed{0.76}$	0.07	0.06	0.01
	Discrimination	-0.01	-0.06	$\boxed{0.24}$	-0.10	-0.05	-0.02
22	Easiness	0.07	0.03	0.07	0.03	$\boxed{0.79}$	0.01
	Discrimination	-0.11	-0.04	-0.10	-0.02	$\boxed{0.31}$	-0.02
23	Easiness	0.06	0.09	0.08	0.08	$\boxed{0.68}$	0.02
	Discrimination	-0.03	-0.13	-0.06	-0.11	$\boxed{0.37}$	-0.04
24	Easiness	0.07	0.08	$\boxed{0.73}$	0.07	0.04	0.01
	Discrimination	-0.02	-0.05	$\boxed{0.19}$	-0.06	-0.04	-0.02
25	Easiness	0.08	0.01	$\boxed{0.67}$	0.05	0.10	0.01
	Discrimination	-0.05	-0.11	$\boxed{0.29}$	-0.04	-0.08	-0.01

26	Easiness	0.06	0.06	0.76	0.07	0.05	0.01
	Discrimination	+0.02	-0.07	0.20	-0.13	-0.02	-0.02
27	Easiness	0.01	0.67	0.08	0.06	0.07	0.02
	Discrimination	-0.01	0.29	-0.06	-0.07	-0.04	-0.04
28	Easiness	0.14	0.04	0.08	0.68	0.05	0.02
	Discrimination	-0.07	-0.05	-0.09	0.67	-0.06	-0.09
29	Easiness	0.07	0.05	0.08	0.05	0.74	0.01
	Discrimination	-0.052	-0.05	-0.13	-0.06	0.27	-0.02
30	Easiness	0.05	0.06	0.09	0.42	0.05	0.01
	Discrimination	-0.05	-0.06	-0.10	0.29	-0.06	-0.02
31	Easiness	0.10	0.75	0.05	0.08	0.02	0.00
	Discrimination	-0.06	0.20	-0.04	-0.08	-0.02	0.00
32	Easiness	0.10	0.66	0.06	0.07	0.10	0.01
	Discrimination	-0.18	0.40	-0.08	-0.07	0.11	-0.03
33	Easiness	0.18	0.11	0.15	0.58	0.06	0.00
	Discrimination	-0.04	-0.06	-0.21	0.28	-0.07	0.00
34	Easiness	0.05	0.08	0.05	0.15	0.66	0.01
	Discrimination	-0.08	-0.06	-0.04	-0.07	0.27	-0.02
35	Easiness	0.10	0.08	0.70	0.06	0.06	0.00
	Discrimination	0.05	-0.12	0.35	-0.06	-0.11	0.00
36	Easiness	0.09	0.75	0.05	0.05	0.06	0.00
	Discrimination	-0.08	0.23	-0.03	-0.06	-0.06	0.00
37	Easiness	0.12	0.65	0.07	0.07	0.08	0.01
	Discrimination	-0.07	0.36	-0.09	-0.06	-0.13	-0.00
38	Easiness	0.62	0.14	0.08	0.06	0.08	0.01
	Discrimination	0.53	-0.21	-0.13	-0.10	-0.07	-0.02
39	Easiness	0.06	0.08	0.64	0.12	0.10	0.00
	Discrimination	-0.04	-0.08	0.39	-0.14	-0.13	0.00
40	Easiness	0.11	0.09	0.06	0.69	0.05	0.00
	Discrimination	-0.11	-0.08	-0.05	0.32	-0.09	0.00
41	Easiness	0.02	0.10	0.57	0.23	0.09	0.01
	Discrimination	-0.02	-0.02	0.20	-0.06	-0.08	-0.02
42	Easiness	0.07	0.71	0.05	0.09	0.08	0.00
	Discrimination	-0.09	0.17	-0.07	-0.09	-0.10	0.00
43	Easiness	0.07	0.06	0.07	0.73	0.05	0.01
	Discrimination	-0.00	-0.11	-0.08	0.26	-0.04	-0.02
44	Easiness	0.19	0.05	0.61	0.06	0.07	0.01
	Discrimination	-0.25	-0.04	0.51	-0.09	-0.09	-0.03

45	Easiness	0.10	0.08	0.05	$\boxed{0.68}$	0.09	0.01
	Discrimination	-0.06	-0.05	-0.04	$\boxed{0.30}$	-0.09	-0.03
46	Easiness	$\boxed{0.67}$	0.04	0.09	0.06	0.15	0.00
	Discrimination	$\boxed{0.38}$	-0.06	-0.12	-0.08	-0.12	0.00
47	Easiness	0.04	0.08	0.08	$\boxed{0.66}$	0.13	0.01
	Discrimination	-0.01	-0.09	-0.05	$\boxed{0.27}$	-0.10	-0.02
48	Easiness	0.14	0.07	0.05	$\boxed{0.64}$	0.10	0.00
	Discrimination	-0.11	-0.05	-0.05	$\boxed{0.33}$	-0.12	0.00
49	Easiness	0.16	0.04	0.05	$\boxed{0.69}$	0.06	0.00
	Discrimination	-0.16	-0.05	-0.04	$\boxed{-0.34}$	-0.09	0.00
50	Easiness	0.12	0.10	0.06	0.06	$\boxed{0.65}$	0.10
	Discrimination	-0.15	-0.07	-0.09	-0.06	$\boxed{-0.40}$	-0.02
51	Easiness	0.05	0.10	0.09	0.06	$\boxed{0.70}$	0.00
	Discrimination	-0.05	-0.04	-0.12	-0.05	$\boxed{0.26}$	0.00
52	Easiness	0.14	0.06	$\boxed{0.70}$	0.07	0.02	0.01
	Discrimination	-0.14	-0.08	$\boxed{0.35}$	-0.09	-0.03	-0.01
53	Easiness	0.07	$\boxed{0.68}$	0.07	0.13	0.05	0.00
	Discrimination	-0.10	$\boxed{0.28}$	-0.04	-0.11	-0.03	0.00
54	Easiness	$\boxed{0.62}$	0.07	0.08	0.08	0.15	0.00
	Discrimination	$\boxed{0.15}$	-0.04	-0.01	-0.04	-0.06	0.00
55	Easiness	0.12	0.09	0.09	0.09	$\boxed{0.60}$	0.01
	Discrimination	-0.05	-0.08	-0.03	-0.06	$\boxed{0.23}$	-0.02
56	Easiness	0.06	0.05	0.04	0.09	$\boxed{0.74}$	0.01
	Discrimination	-0.04	0.07	-0.06	-0.05	$\boxed{0.25}$	-0.03
57	Easiness	0.19	0.06	0.07	$\boxed{0.59}$	0.09	0.00
	Discrimination	-0.04	-0.04	-0.07	$\boxed{0.29}$	-0.14	0.00
58	Easiness	0.05	0.07	0.09	$\boxed{0.70}$	0.09	0.01
	Discrimination	-0.01	-0.10	-0.07	$\boxed{0.34}$	-0.08	-0.01
59	Easiness	$\boxed{0.67}$	0.04	0.08	0.06	0.15	0.00
	Discrimination	$\boxed{0.27}$	-0.05	-0.06	-0.06	-0.11	0.00
60	Easiness	$\boxed{0.71}$	0.06	0.06	0.05	0.11	0.10
	Discrimination	$\boxed{0.37}$	-0.09	-0.09	-0.08	-0.90	-0.20

$\boxed{}$ = Correct alternative

\+ Positive distractor index. Option would to be modified.

APPENDIX IX

Results of Pilot Study

A preliminary analysis of the answer patterns for each of the 81 PAT items showed that:

1. Each item was attempted by majority of the subjects and there was no items that all the subjects answered right or all the subjects answered wrong. This means that all the PAT items satisfied the requirements of the latent trait measurement model.
2. The answered patterns to most of the items were reasonably distributed among the five alternative options. This indicates that the distracters were effective.

The results of the pilot study are organised in two phases: classical test model item analysis and the latent trait test model item analysis.

Phase 1 - Classical test model item analysis

The result of the classical test model item analysis including item difficulties and the point biserial correlation are presented in Table 2.

TABLE 2: Classical Test Model Item analysis

ITEM	Item Score	No. of fail	Item Difficulty	Mean Passes	Point biserial	ITEM	Item Score	No. of fail	Item Difficulty	Mean Passes	Point biserial
1	192	48	0.80	50.77	0.15	41	134	106	0.56	50.57	0.04
2	181	59	0.75	50.82	0.15	42	137	103	0.57	50.79	0.09
3	193	47	0.80	50.83	0.17	43	135	105	0.56	51.28	0.19
4	185	55	0.77	50.86	0.16	44	140	100	0.58	50.89	0.11
5	176	64	0.73	50.49	0.03	45	129	111	0.54	51.00	0.12
6	175	65	0.73	50.81	0.13	46	124	116	0.52	51.01	0.12
7	177	63	0.74	50.71	0.10	47	135	105	0.56	50.41	0.01
8	184	56	0.77	50.89	0.17	48	124	116	0.52	51.40	0.20
9	182	58	0.76	51.09	0.23	49	138	102	0.58	50.57	0.04
10	171	69	0.71	50.88	0.15	50	124	116	0.52	51.10	0.14
11	193	47	0.80	50.80	0.16	51	137	103	0.57	51.11	0.16
12	172	68	0.72	50.63	0.08	52	127	113	0.53	51.59	0.24
13	165	55	0.77	51.17	0.27	53	134	106	0.56	50.83	0.09
14	172	68	0.72	51.27	0.26	54	128	112	0.53	50.79	0.08
15	182	58	0.76	50.98	0.20	55	122	118	0.51	51.16	0.15
16	171	69	0.71	50.95	0.17	56	121	119	0.50	50.38	0.00
17	167	73	0.70	50.87	0.14	57	136	104	0.57	51.32	0.20
18	171	69	0.71	50.82	0.13	58	133	107	0.55	50.95	0.12
19	170	70	0.71	51.23	0.25	59	119	121	0.50	50.45	0.01
20	185	55	0.77	51.31	0.32	60	131	109	0.55	51.24	0.18
21	179	61	0.75	51.01	0.20	61	136	104	0.57	51.24	0.18
22	178	62	0.74	51.13	0.24	62	131	109	0.55	50.89	0.10
23	177	63	0.74	51.04	0.21	63	134	106	0.56	51.13	0.16
24	172	68	0.72	51.24	0.26	64	150	90	0.63	50.34	0.01
25	180	60	0.75	50.92	0.17	65	130	110	0.54	50.62	0.05
26	170	70	0.71	51.09	0.21	66	104	136	0.43	50.69	0.05
27	182	58	0.76	50.99	0.20	67	133	107	0.55	50.81	0.09
28	174	66	0.73	50.91	0.16	68	134	106	0.56	50.98	0.13
29	177	63	0.74	51.23	0.27	69	113	127	0.17	51.11	0.13

ITEM	Item Score	No. of fail	Item Difficulty	Mean Passes	Point biserial
30	172	68	0.72	51.16	0.23
31	190	50	0.79	50.89	0.19
32	202	38	0.84	50.96	0.25
33	158	82	0.66	50.69	0.08
34	155	85	0.65	50.42	0.01
35	146	94	0.61	50.86	0.11
36	159	81	0.66	50.75	0.10
37	133	107	0.55	51.05	0.14
38	134	106	0.56	50.93	0.11
39	127	113	0.53	50.38	0.00
40	129	111	0.54	51.29	0.18

ITEM	Item Score	No. of fail	Item Difficulty	Mean Passe
70	127	113	0.53	51.17
71	113	127	0.47	51.19
72	124	116	0.52	50.82
73	114	126	0.48	51.75
74	124	116	0.52	50.66
75	125	115	0.52	50.93
76	121	119	0.50	51.52
77	121	119	0.50	51.08
78	120	120	0.50	51.07
79	106	134	0.44	50.85
80	136	104	0.57	51.54
81	108	132	0.45	51.36

SUMMARY STATISTICS

	Item Difficulty	Mean Passes	Point biserial
MEAN	0.62	50.96	0.14
STANDARD DEV.	0.11	5.29	0.07

Scores ranged from 35 to 67 (x = 50.96; SD = 5.29). The overall reliability (KR 20 = 0.83).

Item difficulties and point-biserial correlation (validity indices) for each test items was computed and for the total test.

The mean item difficulty was 0.62 (SD = 0.11), ranging from 0.80 to 0.43. The mean point biserial correlation was 0.14 (SD = 0.07), ranging from 0.32 to 0.00.

Phase 2 - The latent Trait Test model item Analysis

The latent Trait Test model item Analysis was performed using the PROX procedures on the Microsoft Excel Visual BASIC computer programme. This programme utilised output from LOGIST to computes several statistics for each item (difficulty) and each person (ability estimate), in addition to summary statistics for all items and ability estimates.

Table 3 shows the estimated item parameter (difficulty) for the revised 81 PAT items. The 81 PAT items ranged in difficulty from -0.44 to 0.41 (X = 0.00; SD = 0.25).

Table 3: Latent Trait Item Model Item Parameter (difficulty) with its Associated Error.

ITEM	Item Score	Difficulty Index	Standard Error	ITEM	Item Score	Difficulty Index	Standard Error
1	192	-0.44	0.16	41	134	0.15	0.13
2	181	-0.30	0.15	42	137	0.12	0.13
3	193	-0.45	0.17	43	135	0.14	0.13
4	185	-0.35	0.16	44	140	0.10	0.13
5	176	-0.24	0.15	45	129	0.19	0.13
6	175	-0.23	0.15	46	124	0.24	0.13
7	177	-0.26	0.15	47	135	0.14	0.13
8	184	-0.34	0.16	48	124	0.24	0.13
9	182	-0.31	0.15	49	138	0.12	0.13
10	171	-0.19	0.14	50	124	0.24	0.13
11	193	-0.45	0.17	51	137	0.12	0.13
12	172	-0.20	0.15	52	127	0.21	0.13
13	165	-0.35	0.16	53	134	0.15	0.13
14	172	-0.20	0.15	54	128	0.20	0.13
15	182	-0.31	0.15	55	122	0.25	0.13
16	171	-0.19	0.14	56	121	0.26	0.13
17	167	-0.15	0.14	57	136	0.13	0.13
18	171	-0.19	0.14	58	133	0.16	0.13
19	170	-0.18	0.14	59	119	0.28	0.13
20	185	-0.35	0.16	60	131	0.18	0.13
21	179	-0.28	0.15	61	136	0.13	0.13
22	178	-0.27	0.15	62	131	0.18	0.13
23	177	-0.26	0.15	63	134	0.15	0.13
24	172	-0.20	0.15	64	150	0.01	0.13
25	180	-0.29	0.15	65	130	0.18	0.13
26	170	-0.18	0.14	66	104	0.41	0.13
27	182	-0.31	0.15	67	133	0.16	0.13
28	174	-0.22	0.15	68	134	0.15	0.13

ITEM	Item Score	Difficulty Index	Standard Error
29	177	-0.26	0.15
30	172	-0.21	0.15
31	190	-0.58	0.16
32	202	-0.58	0.18
33	158	-0.06	0.14
34	155	-0.04	0.14
35	146	0.05	0.13
36	159	-0.07	0.14
37	133	0.16	0.13
38	134	0.15	0.13
39	127	0.21	0.13
40	129	0.19	0.13

ITEM	Item Score	Difficulty Index	Standard Error
69	113	0.33	0.13
70	127	0.21	0.13
71	113	0.33	0.13
72	124	0.24	0.13
73	114	0.32	0.13
74	124	0.24	0.13
75	125	0.23	0.13
76	121	0.26	0.13
77	121	0.26	0.13
78	120	0.27	0.13
79	106	0.39	0.13
80	136	0.13	0.13
81	108	0.37	0.13

SUMMARY STATISTICS

	Difficulty Index	Standard Error
MEAN	0.00	0.14
STANDARD DEV.	0.25	0.01

Item Reliability

One of the main advantages of the latent trait test model over the classical model is that it provides not only an index of reliability for scores on the whole test but also an index of reliability for each person and item.

Each item difficulty or ability estimate is accompanied by its standard error, the most rudimentary indicator of reliability in the classical test model.

Table 3 shows difficulty for each item accompanied by its standard error (reliability) expressed in the same logit units. The standard errors ranged from 0.13 to 0.18 (X = 0.14; SD = 0.01). The mean of the standard errors (0.14) is 0.6 of the standard deviation of the item difficulties. This is a good indication of precision of the test.

In addition to the item difficulties, the latent trait model leads to the assessment of each person ability estimate to the PAT items. Table 4 gives the log ability measure and standard error for each raw score.

Table 4

Log Ability Estimates and Standard Errors for the PAT items

Raw Score	Count	Ability Estimate	Standard Error
35	1	-0.29	1.38
36	1	-0.23	1.40
39	3	-0.08	1.45
40	4	-0.03	1.47
41	3	0.03	1.48
42	2	0.08	1.50
43	5	0.13	1.52
44	11	0.18	1.54
45	13	0.23	1.56
46	16	0.29	1.59
47	15	0.34	1.61
48	18	0.39	1.63
49	14	0.44	1.66
50	17	0.50	1.69
51	17	0.55	1.71
52	20	0.61	1.74
53	16	0.67	1.77
54	11	0.72	1.81
55	12	0.78	1.84
56	13	0.84	1.88
57	8	0.90	1.92

Raw Score	Count	Ability Estimate	Standard Error
58	4	0.96	1.96
59	2	1.03	2.00
60	5	1.10	2.05
61	1	1.16	2.10
62	3	1.23	2.15
63	3	1.31	2.21
64	1	1.38	2.28
67	1	1.63	2.51
68	0	1.73	2.60
69	0	1.82	2.71
70	0	1.93	2.97
71	0	2.04	3.13
72	0	2.17	3.13
73	0	2.31	3.32
74	0	2.46	3.55
75	0	2.64	3.83
76	0	2.84	4.20
77	0	3.09	4.69
78	0	3.40	5.42
79	0	3.84	6.64
80	0	4.57	9.39
81	0		

Total	240		

Furthermore, the chi-square fit statistic was computed for each of the 81 PAT items.

Table 5 shows the proportion of subjects answering an item correctly compared to the proportions expected to answer an item correctly.

The chi-square test of fit involves a division of the sample into ability subgroups by score levels. In the present analysis, six groups were used. The mean difficulty, standard deviations and the group counts appear below the table. The overall fit was 83.36 (df = 400, N=240) and was statistically significant at 0.01 level of significance. The associated degree of freedom was (k-1)(r-1).

k = number of score groups = 6

r = number of items = 81

Table 5

Analysis of fit

	Observed proportions of subjects answering an item correctly						Expected proportions of subjects to answ an item correctly					
Item	GRP 1 61-67	GRP 2 56-60	GRP 3 51-55	GRP 4 46-50	GRP 5 41-45	GRP 6 35-40	GRP 1 61-67	GRP 2 56-60	GRP 3 51-55	GRP 4 46-50	GRP 5 41-45	GR 35-
1	0.89	0.88	0.83	0.78	0.73	0.67	0.79	0.77	0.69	0.60	0.54	0.45
2	0.89	0.94	0.76	0.65	0.73	0.77	0.79	0.88	0.60	0.42	0.54	0.60
3	0.78	0.81	0.90	0.83	0.62	0.56	0.61	0.66	0.80	0.68	0.38	0.31
4	1.00	0.91	0.78	0.71	0.71	0.78	1.00	0.82	0.60	0.51	0.50	0.61
5	0.88	0.69	0.71	0.80	0.68	0.57	0.78	0.47	0.50	0.64	0.46	0.32
6	0.89	0.91	0.70	0.70	0.65	0.77	0.79	0.82	0.49	0.49	0.42	0.60
7	1.00	0.72	0.74	0.76	0.68	0.54	1.00	0.52	0.54	0.58	0.46	0.30
8	0.78	0.81	0.84	0.75	0.68	0.44	0.61	0.66	0.71	0.56	0.46	0.20
9	0.89	0.82	0.88	0.72	0.50	0.67	0.79	0.66	0.78	0.52	0.25	0.45
10	0.89	0.72	0.79	0.68	0.59	0.86	0.79	0.52	0.62	0.40	0.30	0.44
11	0.78	0.91	0.86	0.78	0.85	0.89	0.61	0.82	0.73	0.60	0.42	0.79
12	0.66	0.75	0.79	0.68	0.71	0.45	0.44	0.56	0.62	0.46	0.50	0.20
13	0.89	0.91	0.86	0.73	0.62	0.45	0.79	0.82	0.73	0.53	0.38	0.20
14	0.89	0.91	0.76	0.68	0.59	0.34	0.79	0.82	0.58	0.46	0.34	0.11
15	1.00	0.81	0.81	0.72	0.68	0.45	1.00	0.66	0.66	0.53	0.46	0.20
16	0.89	0.75	0.82	0.61	0.74	0.34	0.79	0.56	0.67	0.37	0.54	0.11
17	0.89	0.75	0.74	0.63	0.68	0.67	0.79	0.56	0.54	0.39	0.46	0.45
18	0.78	0.85	0.71	0.70	0.68	0.44	0.61	0.71	0.50	0.49	0.46	0.20
19	0.89	0.91	0.75	0.66	0.56	0.45	0.79	0.82	0.56	0.44	0.31	0.20
20	1.00	0.91	0.89	0.71	0.50	0.57	1.00	0.82	0.80	0.51	0.25	0.32
21	1.00	0.81	0.80	0.69	0.68	0.56	1.00	0.66	0.65	0.47	0.46	0.31
22	1.00	0.84	0.78	0.75	0.54	0.54	1.00	0.71	0.60	0.56	0.28	0.30
23	1.00	0.85	0.79	0.89	0.59	0.67	1.00	0.71	0.62	0.47	0.35	0.45
24	1.00	0.94	0.71	0.68	0.82	0.45	1.00	0.88	0.50	0.46	0.38	0.20
25	0.89	0.87	0.78	0.69	0.74	0.56	0.79	0.76	0.60	0.47	0.54	0.3[illegible]
26	0.89	0.78	0.74	0.73	0.56	0.43	0.79	0.61	0.54	0.53	0.31	0.1[illegible]
27	0.89	0.91	0.76	0.79	0.65	0.57	0.79	0.82	0.58	0.62	0.31	0.3[illegible]

	Observed proportions of subjects answering an item correctly						Expected proportions of subjects to answer an item correctly					
28	0.89	0.72	0.84	0.64	0.65	0.68	0.79	0.52	0.71	0.41	0.42	0.46
29	1.00	0.88	0.77	0.73	0.53	0.66	1.00	0.77	0.58	0.53	0.28	0.44
30	0.78	0.94	0.76	0.65	0.59	0.55	0.61	0.88	0.58	0.42	0.35	0.31
31	0.89	0.91	0.78	0.83	0.68	0.55	0.79	0.82	0.60	0.68	0.46	0.31
32	1.00	0.91	0.93	0.78	0.74	0.67	1.00	0.82	0.87	0.60	0.54	0.45
33	0.77	0.68	0.71	0.61	0.62	0.90	0.60	0.47	0.50	0.38	0.38	0.31
34	0.56	0.54	0.74	0.63	0.64	0.56	0.31	0.28	0.54	0.39	0.42	0.31
35	1.00	0.69	0.57	0.59	0.59	0.55	1.00	0.47	0.32	0.34	0.35	0.30
36	0.78	0.69	0.70	0.64	0.62	0.57	0.61	0.47	0.49	0.41	0.38	0.31
37	0.67	0.54	0.61	0.59	0.41	0.33	0.45	0.28	0.37	0.35	0.17	0.11
38	0.89	0.62	0.59	0.47	0.50	0.67	0.79	0.39	0.35	0.23	0.25	0.45
39	0.22	0.50	0.63	0.49	0.56	0.34	0.05	0.25	0.40	0.24	0.31	0.11
40	0.78	0.66	0.54	0.54	0.42	0.34	0.61	0.43	0.29	0.29	0.17	0.11
41	0.67	0.47	0.62	0.59	0.44	0.45	0.45	0.22	0.38	0.34	0.19	0.20
42	0.45	0.72	0.57	0.54	0.56	0.55	0.20	0.52	0.32	0.29	0.31	0.30
43	0.89	0.72	0.61	0.44	0.56	0.45	0.79	0.51	0.37	0.19	0.31	0.20
44	0.67	0.72	0.55	0.60	0.50	0.44	0.44	0.52	0.31	0.36	0.25	0.20
45	0.67	0.56	0.63	0.47	0.50	0.23	0.45	0.32	0.40	0.22	0.25	0.05
46	0.67	0.53	0.55	0.50	0.50	0.23	0.45	0.28	0.30	0.25	0.25	0.05
47	0.67	0.62	0.58	0.46	0.62	0.77	0.45	0.39	0.33	0.21	0.38	0.60
48	0.55	0.81	0.49	0.51	0.32	0.45	0.31	0.66	0.24	0.26	0.10	0.20
49	0.89	0.53	0.63	0.47	0.68	0.44	0.79	0.28	0.40	0.22	0.46	0.20
50	0.67	0.53	0.60	0.45	0.50	0.33	0.45	0.28	0.35	0.20	0.25	0.11
51	0.89	0.78	0.53	0.54	0.52	0.32	0.79	0.61	0.28	0.29	0.28	0.11
52	0.45	0.76	0.61	0.51	0.26	0.34	0.20	0.57	0.37	0.26	0.07	0.11
53	0.88	0.62	0.54	0.53	0.56	0.45	0.78	0.39	0.29	0.28	0.31	0.20
54	0.78	0.44	0.65	0.47	0.41	0.68	0.61	0.19	0.42	0.22	0.17	0.46
55	0.78	0.69	0.47	0.49	0.41	0.44	0.61	0.47	0.22	0.24	0.17	0.20
56	0.44	0.50	0.54	0.51	0.41	0.55	0.20	0.25	0.29	0.26	0.17	0.31
57	0.89	0.69	0.57	0.55	0.47	0.34	0.79	0.47	0.32	0.30	0.22	0.11
58	0.89	0.53	0.63	0.46	0.53	0.56	0.79	0.28	0.40	0.21	0.28	0.31

	Observed proportions of subjects answering an item correctly						Expected proportions of subjects to answer an item correctly					
59	0.44	0.56	0.47	0.54	0.41	0.45	0.20	0.32	0.22	0.29	0.17	0.
60	0.78	0.65	0.59	0.51	0.41	0.33	0.61	0.43	0.35	0.26	0.17	0.
61	0.89	0.66	0.57	0.56	0.50	0.22	0.79	0.43	0.32	0.32	0.25	0.
62	0.67	0.56	0.59	0.56	0.41	0.33	0.45	0.31	0.35	0.32	0.17	0.
63	0.78	0.66	0.63	0.45	0.55	0.32	0.60	0.43	0.40	0.20	0.31	0.
64	0.89	0.53	0.62	0.66	0.55	0.66	0.79	0.28	0.38	0.41	0.31	0.
65	0.55	0.63	0.55	0.52	0.50	0.46	0.31	0.39	0.31	0.27	0.25	0.
66	0.45	0.53	0.45	0.36	0.44	0.55	0.20	0.28	0.20	0.13	0.20	0.
67	0.56	0.63	0.54	0.61	0.41	0.43	0.31	0.39	0.29	0.38	0.17	0.
68	0.88	0.59	0.54	0.59	0.53	0.11	0.78	0.35	0.29	0.34	0.28	0.
69	0.78	0.60	0.38	0.50	0.44	0.34	0.61	0.35	0.15	0.25	0.20	0
70	0.55	0.75	0.58	0.41	0.50	0.43	0.30	0.56	0.34	0.17	0.25	0.
71	1.00	0.50	0.46	0.44	0.41	0.45	1.00	0.25	0.21	0.19	0.17	0
72	0.67	0.57	0.50	0.50	0.50	0.57	0.45	0.32	0.25	0.25	0.25	0
73	0.67	0.75	0.46	0.48	0.27	0.22	0.44	0.56	0.21	0.23	0.07	0
74	0.33	0.66	0.50	0.50	0.55	0.22	0.11	0.43	0.25	0.25	0.31	0
75	0.67	0.63	0.52	0.49	0.47	0.44	0.44	0.39	0.28	0.24	0.22	0
76	1.00	0.62	0.45	0.54	0.33	0.33	1.00	0.39	0.21	0.29	0.11	0
77	0.55	0.60	0.58	0.40	0.53	0.33	0.30	0.35	0.34	0.16	0.28	0
78	0.67	0.69	0.42	0.56	0.38	0.23	0.45	0.47	0.18	0.32	0.14	0
79	0.56	0.44	0.50	0.40	0.38	0.44	0.31	0.19	0.25	0.16	0.15	0
80	1.00	0.84	0.52	0.54	0.44	0.33	1.00	0.71	0.26	0.29	0.19	0
81	0.44	0.59	0.54	0.36	0.41	0.11	0.20	0.35	0.29	0.13	0.17	0

	GRP 1	GRP 2	GRP 3	GRP 4	GRP 5	GRP 6
Group Count	9	32	76	80	34	9
Mean Difficulty	0.78	0.71	0.65	0.59	0.54	0.48
Standard Deviation	0.18	0.14	0.13	0.12	0.12	0.16

Total Chi-Square Value 83.36
Degree of Freedom 400
Significant at $\leq$ 0.01 level