DEVELOPMENT AND VALIDATION OF A TEST OF INTEGRATED SCIENCE PROCESS SKILLS FOR THE FURTHER EDUCATION AND TRAINING LEARNERS

By

KAZENI MUNGANDI MONDE MONICA

A DISSERTATION SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF SCIENCE IN SCIENCE EDUCATION

IN THE FACULTY OF NATURAL AND AGRICULTURAL SCIENCES
UNIVERSITY OF PRETORIA
SOUTH AFRICA

2005

Declaration

I hereby declare that this dissertation is the result of my own investigations and has not been previously submitted for any degree or diploma in any University. To the best of my knowledge, this dissertation contains no materials previously published by any other person, except where acknowledged.

Signature: ……..

M.M.M.Kazeni

Date: ……

Dedication

This dissertation is dedicated to my late mother, Mulai, Mushele Mungandi, and my father, Joseph Mungandi.

ACKNOWLEDGEMENTS

I am grateful to the Almighty God for His protection, guidance, providence, and especially for sparing my life. Glory be to God, for without Him, this study would have been futile.

I would like to thank my supervisor, Professor G. O. M. Onwu, for suggesting the research topic, and for his guidance and constructive criticism throughout the course of the study. I would also like to acknowledge the financial contribution made by the National Research Fund (NRF), as a student research grant linked bursary, which was made on the recommendation of Professor G.O.M. Onwu.

I would also like to earnestly and gratefully thank my dear friend Dr. B.S. Linyama for his intellectual, moral and financial support at every stage of this study. I owe the completion of this study to him.

Special thanks to the UNIFY staff, who supported and advised me on various aspects of the study. Special thanks to Mrs. E. Malatjie, Mr. S. S. Mathabatha, Mrs P. C. Mathobela, Dr. K. M. Chuene and Mr. M. T. Mabila.

Lastly, but not the least, I would like to thank my children, Mulai and Chudwa for allowing me to concentrate on the study at the expense of their well being.

ABSTRACT

The South African Revised National Curriculum Statement (RNCS), curriculum guides, and instructional materials on the Outcomes Based Education (OBE), emphasize the development and use of science process skills. Learners using these materials are expected to acquire these skills. The traditional assessment of process skills through practical work only, has practical constraints, particularly in large under resourced classes. A reliable, convenient and cost effective complementary paper and pencil test for assessing these skills may provide a solution. In South Africa, little research has been undertaken in the area of development and validation of science process skills tests. This study was an attempt to develop and validate a test of integrated science process skills, referenced to a specific set of objectives, for use in the further education and training band (grades 10 – 12). The science process skills tested for were: identifying and controlling variables, stating hypotheses, experimental design, graphing and interpreting data, and operational definitions. Thirty multiple-choice items, designed to be content independent; and gender, race, school type, and location neutral, were developed and administered to a total of 1043 grade 9, 10, and 11 learners from ten schools, in the Limpopo province of South Africa. Results from data analysis show that the test is valid, and that its test characteristics fall within the acceptable range of values for discrimination index, index of difficulty, reliability, and readability levels. Comparison of the performance of different groups of learners who wrote the test showed that the test is gender and race neutral.

CERTIFICATION BY SUPERVISOR

I certify that this work was carried out by Kazeni – Mungandi Monde Monica of the Joint Centre for Science, Mathematics and Technology Education, Faculty of Natural and Agricultural Sciences, at the University of Pretoria, Pretoria, South Africa.

Supervisor _____

Prof. G.O.M. Onwu
Department of Science, Mathematics and Technology Education
Groenkloof campus
University of Pretoria
Pretoria
Republic of South Africa.

Date: _____

TABLE OF CONTENTS. **PAGE**

LIST OF TABLES

LIST OF GRAPHS

CHAPTER 1
INTRODUCTION

This chapter outlines the background and rationale of the study, the research questions, the significance of the study, its scope, and the basic assumptions made in the study.

1.1 BACKGROUND AND RATIONALE OF THE STUDY

During the 1960s and 70s, science curriculum innovations and reforms were characterised by attempts to incorporate more inquiry oriented and investigative activities into science classes (Dillashaw and Okey, 1980). Teachers attempted to move their students into the world of science, especially the world of research scientists. This involved consideration of the processes used by such scientist and the concepts they used. These moves were also accompanied by similar efforts to measure the outcomes of such processes (Onwu and Mozube, 1992).

In the new South Africa, the government's realization of its inheritance of an inefficient and a fragmented educational system, the formulation of the critical outcomes of education, and to some extent the poor performance of South African learners in the Third International Mathematics and Science Study (TIMSS) results (HSRC, 2005b; Howie, 2001) revealed a deficiency in Science education. Several papers, including government white papers were published (National department of Education, 1996; 1995), which attempted to address the deficiencies, and shape the educational policies in the country (Howie, 2001).

The publication of the South African National Qualifications framework, as well as policy guidelines have provided the blue print for change and reform, and once implemented should significantly improve the quality of education offered in accordance with the principles of the Outcomes Based Education (OBE) of the new Curriculum 2005 (Onwu and Mogari, 2004; Department of Education, 1996; 1995).

The Natural Science learning area of the Curriculum 2005 emphasizes the teaching and learning of science process skills (Department of Education, 2002). The South African Revised National Curriculum Statement (RNCS) refers to science process skills as: "The learner's cognitive ability of creating meaning and structure from new information and experiences" (Department of education, 2002, pp13). The emphasis placed on the development and use of science process skills by the Revised National Curriculum Statement is evident when it expresses the view that from a teaching point, process skills are the building blocks from which suitable science tasks are constructed. It further argues that a framework of process skills enables teachers to design questions, which promote the kind of critical thinking required by the curriculum 2005 Learning Outcomes (Department of Education 2002).

From a learning point of view, process skills are an important and necessary means by which the learner engages with the world and gains intellectual control of it through the formation of concepts and development of scientific thinking (Department of Education 2002).

The scientific method, scientific thinking, and critical thinking have been terms used at various times to describe these science skills. However, as Padilla (1990) has noted, the use of the term 'science process skills' in place of those terms was popularised by the curriculum project, Science A Process Approach (SAPA). According to SAPA, science process skills are defined as a set of broadly transferable abilities appropriate to many science disciplines and reflective of the behaviour of scientists (Padilla, 1990).

Science process skills are hierarchically organized, ranging from the simplest to the more complex ones. This hierarchy has been broadly divided into two categories, namely, the primary (basic) science process skills, and the integrated (higher order) science process skills (Dillashaw and Okey, 1980; Padilla, 1990; The American Association for the Advancement of Science-AAAS, 1998). Integrated science process skills are science process skills that incorporate (integrate) or involve the use of different basic science process skills, which provide a foundation for learning the more complex (integrated)

science skills (Rezba, Sparague, Fiel, Funk, Okey, and Jaus, 1995). The ability to use the integrated science process skills is therefore dependent on the knowledge of the simpler primary (basic) processes, (Onwu and Mozube, 1992). Integrated science process skills are high order thinking skills which are usually used by scientists when designing and conducting investigations (Rezba, *et al*. 1995). This study deals with the assessment of the integrated science process skills.

The Revised National Curriculum Statement identifies several science process skills as being essential in creating outcomes-based science tasks (Department of Education, 2002). These science process skills are incorporated in all the three science learning areas (scientific investigations, constructing science knowledge, and science, society and the environment) of the science curriculum 2005 (Department of Education, 2002). In consequence, many of the science curriculum guides and instructional materials of the new Outcomes Based Education have, as important outcomes, the development of Science Process Skills. Learners using these instructional materials are expected to acquire science process skills, such as; formulating hypotheses, identifying, controlling and manipulating variables, operationally defining variables, designing and conducting experiments, collecting and interpreting data, and problem solving, in addition to mastering the content of the subject matter (Department of education, 2002).

Having established the importance that is attached to science process skills by the Revised National curriculum statement, the question that arises is; to what extent have the learners who use this curriculum and the related instructional materials acquired the science process skills? The answer to this question lies in the effective assessment of learners' competence in those specific skills. A review of the literature in the South African setting shows that not much work, if any at all, has been done in the area of test construction and validation, for use to assess these specific skills, especially for the Further Education and Training (FET) band. The search for available literature on science process skills in the South African setting showed the need for the development of a test geared towards the FET natural science learners.

The traditional methods of assessing science process skills competence, such as through practical work only, has a number of practical constraints, particularly in the context of teaching and learning in large under-resourced science classes (Onwu, 1999; 1998). First, most South African schools, especially those in rural areas are characterised by large science classes (Flier, Thijs and Zaaiman, 2003; Human Sciences Research Council-HSRC, 2005a, 1997), which are difficult to cater for during practicals. Second, most of these schools either do not have laboratories, or have poorly equipped ones (Muwanga-Zake, 2001a). This makes expectations of effective practical work unrealistic. Thirdly, many science classes in South Africa are taught by either unqualified or under qualified science educators (Arnott, Kubeka, Rice & Hall, 1997; Human Sciences Research Council-HSRC, 1997). These educators may not be competent to teach and assess inquiry science (use of science process skills) through practical work, because of their lack of familiarity with science processes and apparatus. This may undermine their practical approach to science (Muwanga-Zake, 2001a), and resort to a theoretical one. Science educators in South Africa, therefore need a convenient and cost effective means of assessing science process skills competence effectively and objectively.

It is true that a hands on activity procedure would seem most appropriate for assessing process skills competence, but as indicated, the stated constraints pose enormous practical assessment problems in the South African context. As a result, it became necessary to seek alternative ways of assessing learners' competence in these skills. Hence the need for this study, which attempted to develop and validate a science process skills test, which favours no particular science discipline, for use with FET learners.

One of the ways that have been used to assess science process skills, especially in large under-resourced science classes is through the use of paper and pencil format, which does not require expensive resources (Onwu and Mozube, 1992; Tobin and Capie, 1982; Dillashaw and Okey, 1980). There are a number of paper and pencil science process skills tests in existence, but most of these tests would appear to present some challenges that are likely to make them unsuitable for use in the South African context.

The main challenge is that most of these tests have been developed and validated outside South Africa. As a result, adopting the existing tests is likely to be problematic. First, the language used in most of the tests does not take cognisance of second and third English language users. In consequence, most of the examples and terminologies used in the tests may be unfamiliar to most South African learners who use English as a second or even third language. Second, the tests also contain a lot of technical or scientific terms in their text that may not be familiar to novice science learners. Examples of such technical terms include; hypothesis, variables, manipulated variables, operational definition (eg. in TIPS II, by Burns, *et al*, 1985, and TIPS, by Dillashaw and Okey, 1980).

Thirdly, research has shown that learners learn process skills better if they are considered an important object of instruction relatable to their environment, using proven teaching methods (Magagula and Mazibuko, 2004; Muwanga-Zake, 2001b). In other words, the development and acquisition of skills is contextual. Other researchers have raised concerns regarding the exclusive use of unfamiliar materials or conceptual models in African educational systems (Magagula and Mazibuko, 2004; Okrah, 2004, 2003, 2002; Pollitt and Ahmed, 2001). These researchers advocate for the use of locally developed educational materials that are familiar, and which meet the expectations of the learners. The use of the foreign developed existing tests with South African learners may therefore lead to invalid results (Brescia and Fortune, 1988).

1.2 THE PURPOSE OF THE STUDY

The purpose of this study was to develop and validate, a reliable, convenient and cost effective paper and pencil test, for measuring integrated science process skills competence effectively and objectively in the natural sciences further education and training band, and which favours no particular subject discipline, school type, gender, location, or race.

1.3. RESEARCH QUESTIONS

In order to achieve the purpose of this study, the following research questions were addressed;

1. Is the developed test instrument a valid and reliable means of measuring learners' competence in Integrated Science Process Skills, in terms of its test characteristics?
2. Does the developed test instrument show sensitivity in regard to learners of different races, gender, school type, and location, as prevalent in the South African context?

1.3.1 OBJECTIVES OF THE STUDY.

The determination of the validity and reliability of a test instrument involves the estimation of its test characteristics, which should fall within the accepted range of values. One way to assure that a test is sensitive (un-biased) towards different groups of participants is to build fairness into the development, administration, and scoring processes of the test (Zieky, 2002). In order to fulfil these requirements, the following objectives were set for the study.

1. To develop a paper and pencil test of integrated science process skills, referenced to a specific set of objectives for each skill.
2. To construct test items that fall within the accepted range of values for reliable tests in each of the test characteristics of; validity, reliability, item discrimination index, index of difficulty, and readability level.
3. To construct test items that do not favour any particular science discipline, or participants belonging to different school types, gender, race, or location.
4. To construct test items that do not contain technical and unfamiliar terminology.

1.4 SIGNIFICANCE OF THE STUDY

While policies, content, learning outcomes, assessment standards and teaching instructions are meticulously prescribed in the Revised National Curriculum Statement (Department of education, 2002), the responsibility of assessing the acquisition of higher order thinking skills and the achievement of the prescribed assessment standards lie with the educators. There seems to be a policy void on how to address the constraints related to the assessment of science process skills in large under-resourced science classes. Educators therefore use different assessment methods and instruments of varying levels and quality, to accomplish the task of assessing those skills. In most cases, educators use un-validated, unreliable and biased assessment tools, because of the many hurdles associated with the assessment of science process skills (Muwanga-Zake, 2001a; Berry, Mulhall, Loughran and Gunstone, 1999; Novak and Govin, 1984; Dillashaw and Okey, 1980).

This study is significant in that, the developed instrument is an educational product that is developed and validated within the South African context. It is hoped that it will provide teachers with a valid and reliable cost effective means of measuring science process skills attainment effectively and objectively. The developed test is likely to provide a useful practical solution to the problem of assessing science process skills in large under-resourced science classes.

Furthermore, researchers who may want to identify the process skills inherent in certain curricula material, determine the level of acquisition of science process skills in a particular unit, or establish science process skills competence by science teachers, need a valid, reliable, convenient, efficient, and cost-effective assessment instrument to work with. It is hoped that the developed instrument could be used for this purpose. It is also envisaged that researchers would use the procedure used in the development of this test to develop and validate similar assessment instruments.

The developed test could be used for baseline, diagnostic, or formative assessment purposes, especially by those teaching poorly resourced large classes, as it does not require expensive resources. Moreover, as a locally developed test, it will be readily available to South African educators, together with its marking key.

Lastly, the attempt to make the developed test gender, racial, and location sensitive will provide a neutral (un-biased) assessment instrument for the test users, in terms of their ability to demonstrate competence in the integrated science process skills.

1.5 THE SCOPE OF THE STUDY

As indicated in section 1.2, this study was concerned with the development and validation of a test of integrated (higher order) science process skills only. The specific skills considered are, identifying and controlling variables, stating hypotheses, operational definitions, graphing and interpreting data, and experimental design. In the South African context, these high order thinking skills are learned with sustained rigor at the Further Education and Training band –FET, (grades 10 –12)] (Department of Education, 2002). This study therefore involved learners from the FET band.

The study was undertaken based on the assumptions that, the learners from the schools that participated in the study have been using the Revised National Curriculum Statement, which emphasizes the teaching and learning of science process skills, and that learners of the same grade who participated in the study had covered the same syllabus.

1.6 OVERVIEW OF THE STUDY REPORT

The first chapter discusses the rationale and purpose of the study, the research questions and objectives, its significance, and its scope. The second chapter reviews and discusses literature that is relevant to the study. This review includes a discourse on the conceptual

framework of the study, existing research on science process skills development and academic ability, development of science process skills tests outside South Africa, development of science process skills in South Africa, and an overview of the criteria for test development and validation. The third chapter outlines the methodology of the study. It describes the research design, population and sample description, instrumentation, pilot study, the main study, statistical procedures used in the main study, and ethical issues. The fourth chapter provides an analysis and discussion of the findings of the study. The fifth chapter summarises the results and draws conclusions from them. It also discusses the educational implications of the study, and recommendations based on the study. The chapter ends with a discussion of the limitations of the study and areas for further research. The reference and appendices follow chapter five.

CHAPTER 2
LITERATURE REVIEW.

This chapter reviews the literature that relates to the development and validation of science process skills tests. The review is organised under the following sub-headings; the conceptual framework of the study, science process skills development and academic abilities, the development of science process skills tests outside South Africa, tests developed for primary school level, tests developed for secondary school level, development of science process skills tests in South Africa, the criteria used for test development and validation.

2.1 CONCEPTUAL FRAMEWORK OF THE STUDY

In our increasingly complex and specialized society, it is becoming imperative that individuals are capable of thinking creatively, critically, and constructively. These attributes constitute higher order thinking skills (Wiederhold, 1997). Nitko (1996) included the ability to use reference material, and interpret graphs, tables and maps among the high order thinking skills. Thomas and Albee (1998) defined higher order thinking skills as thinking that takes place in the higher levels of the hierarchy of cognitive processing. The concept of higher order thinking skills became a major educational agenda item with the publications of Bloom's taxonomy of educational objectives (Bloom, Englehart, Furst and Krathwohl 1956). Bloom and his co-workers established a hierarchy of educational objectives, which attempts to divide cognitive objectives into subdivisions, ranging from the simplest intellectual behaviour to the most complex ones. These subdivisions are: knowledge, comprehension, application, analysis, synthesis, and evaluation (Wiederhold, 1997). Of these objectives, application, analysis, synthesis, and evaluation are considered to be higher order thinking skills (Wiederhold, 1997).

Demonstration of competence in integrated science process skills is said to require the use of higher order thinking skills, since competence in science process skills entails the ability to apply learnt material to new and concrete situations, analyse relationships between parts and the recognition of the organizational principles involved, synthesize parts together to form a new whole, and to evaluate or judge the value of materials, such as, judging the adequacy with which conclusions are supported by data (Baird and Borick, 1985). Nonetheless, different scholars interpret performance on tests of higher order thinking or cognitive skills differently, because there are no agreed-upon operational definitions of those skills. Developing such definitions is difficult because our understanding of process skills is limited. For example, we know little about the relationship between low order thinking skills and higher order thinking skills. Improved construction and assessment of higher order cognitive skills is contingent on developing operational definitions of those skills. The many theoretical issues surrounding the relationship between discipline knowledge and cognitive skills are by no means resolved.

In spite of this limited understanding of cognitive skills, most work in the cognitive psychology suggest that use of higher order cognitive skills is closely linked with discipline specific knowledge. This conclusion is based primarily on research in Problem solving and learning to learn skills (Novak and Govin, 1984). As it is, the conclusion is limited to these specific higher order thinking skills, and may be different for higher order thinking skills such as inductive or deductive reasoning.

The close relationship between science process skills and higher order thinking skills is acknowledged by several researchers. For instance, Padilla, *et al* (1981) in their study of "The Relationship between Science Process Skills and Formal Thinking Abilities," found that, formal thinking and process skills abilities are highly inter-related. Furthermore, Baird and Borick (1985), in their study entitled "Validity Considerations for the Study of Formal Reasoning and Integrated Science Process Skills", concluded that, Formal Reasoning and Integrated Science Process Skills competence share more variance than expected, and that they may not comprise distinctly different traits.

The format for assessing integrated science process skills is based on that of assessing higher order thinking skills, as indicated by Nitko (1996), who contends that the basic rule for crafting assessment of higher order thinking skills is to set tasks requiring learners to use knowledge and skills in novel situations. He asserts that assessing higher order thinking skills requires using introductory material as a premise for the construction of the assessment task(s). He cautions that, to assess high order thinking skills, one should not ask learners to simply repeat the reasons, explanation or interpretations they have been taught or read from some source (Nitko, 1996), but that tasks or test items should be crafted in such a way that learners must analyse and process the information in the introductory material to be able to answer the questions, solve the problems or otherwise complete the assessment tasks. The format used in the development of test items for assessing integrated science process skills in this study was based on the above stated principles. This study was therefore guided by the conceptual framework of the assessment of higher order thinking skills.

2.2 SCIENCE PROCESS SKILLS DEVELOPMENT AND ACADEMIC ABILITY

Given the emphasis placed on the development and use of science process skills by the South African Revised National Curriculum Statement, the question that comes to one's mind is, what is the relationship between science process skills development and academic ability? Research has highlighted the relevance of science process skills development on academic ability.

First, it should be noted that what we know about the physical world today is a result of investigations made by scientists in the past. Years of practice and experience have evolved into a particular way of thinking and acting in the scientific world. Science process skills are the 'tools' scientists use to learn more about our world (Osborne and Fryberg, 1985; Ostlund, 1998). If learners have to be the future scientists, they need to learn the values and methods of science. The development of science process skills is

said to empower learners with the ability and confidence to solve problems in every day life.

Secondly, research literature shows that science process skills are part of and central to other disciplines. The integration of science process skills with other disciplines has produced positive effects on student learning. For instance, Shann (1977) found that teaching science process skills enhances problem-solving skills in mathematics. Other researchers found that science process skills not only enhance the operational abilities of kindergarten and first grade learners, but also facilitate the transition from one level of cognitive development to the next, among older learners (Froit, 1976; Tipps, 1982). Simon and Zimmerman (1990) also found that teaching science process skills enhances oral and communication skills of students. These researchers agree with Bredderman's (1983) findings in his study of the effect of activity based elementary science on student outcomes, that the process approach programmes of the sixties and seventies, such as the Elementary Science Study (ESS), Science Curriculum Improvement Study (SCIS) and Science-A Process Approach (SAPA), were more effective in raising students' performance and attitudes than the traditional based programmes.

Ostlund (1998) pointed out that the development of scientific processes simultaneously develops reading processes. Harlen (1999) reiterated this notion by stating that science processes have a key role to play in the development of skills of communication, critical thinking, problem solving, and the ability to use and evaluate evidence. Competence in science process skills enables learners to learn with understanding. According to Harlen, learning with understanding involves linking new experiences to previous ones, and extending ideas and concepts to include a progressively wider range of related phenomena. The role of science process skills in the development of 'learning with understanding' is of crucial importance. If science process skills are not well developed, then emerging concepts will not help in the understanding of the world around us (Harlen, 1999). Harlen suggested that science process skills should be a major goal of science education because science education requires learners to learn with understanding.

Having established the positive effects of science process skills on learners' academic abilities, the need to assess the development and achievement of these important outcomes (science process skills) becomes imperative. Harlen (1999) emphasized the need to include science process skills in the assessment of learning in Science. She contends that without the inclusion of science process skills in science assessment, there will continue to be a mismatch between what our students need from Science, and what is taught and assessed (Harlen, 1999). She further argued that assessing science process skills is important for formative, summative and monitoring purposes because the mental and physical skills described as science process skills have a central part in learning with understanding.

Unfortunately, the assessment of the acquisition of these important skills is still not a routine part of the evaluation process in educational systems, including the South African educational system.

Some critics have urged against the effectiveness of science process skills in enhancing academic ability (Gott, R. and Duggan, S. 1996; Millar, R., Lubben, F., Gott, R. and Duggan, S. 1994; Millar and Driver, R. 1987). These researchers have questioned the influence of science process skills on learner performance, and their role in the understanding of evidence in Science. Millar and Driver, R. (1987) present a powerful critique on the independence of science process skills from content. They argue that science process skills can not exist on their own without being related to content. This argument is valid. However, content independence in the context of this study does not mean that the items are completely free from content, it rather means that the student does not require in-depth knowledge of the content (subject) to be able to demonstrate the required science process skill. Some researchers have generally criticized the positivist approach to measurement. While it is acknowledged that these critics present valid and compelling arguments against the use of positivist approach to measurement, and the effectiveness of science process skills in enhancing ability, the evidence regarding their success is overwhelming, as reviewed above. I personally appreciate the issues raised against the effective use of science process skills, but I am of the opinion that they play a

vital role in the understanding of science as a subject, as well as the acquisition of Science skills necessary for everyday survival.

2.3 DEVELOPMENT OF SCIENCE PROCESS SKILLS TESTS OUTSIDE SOUTH AFRICA

The educational reforms of the 60s and 70s prompted the need to develop various instruments for testing the acquisition of science process skills (Dillashaw and Okey, 1980). Several researchers developed instruments to measure the process skills that are associated with inquiry and investigative abilities, as defined by Science – A Process Approach (SAPA), and the Science Curriculum Improvement Study (SCIS) (Dillashaw and Okey, 1980). There were efforts to develop science process skills tests for both primary and secondary school learners. The literature on the development of science process skills tests for the different levels of education, show some shortcomings that prompted subsequent researchers to develop more tests in an attempt to address the identified shortcomings.

2.3.1 TEST DEVELOPMENT FOR PRIMARY SCHOOL LEVEL

The researchers who developed the early science process skills tests for primary school learners include: Walbesser (1965), who developed a test of basic and integrated process skills, especially intended for elementary children using the SAPA curriculum program. Dietz and George (1970) used multiple-choice questions to test the problem solving skills of elementary students. This test established the use of written tests as a means to measure problem-solving skills (Lavinghousez,1973). In 1972, Riley developed the test of science inquiry skills for grade five students, which measured the science process skills of identifying and controlling variables, predicting and inferring, and interpreting data, as defined by SCIS (Dillashaw and Okey, 1980). McLeod, Berkheimer, Fyffe, and Robison (1975) developed the group test of four processes, to measure the skills of controlling variables, interpreting data, formulating hypotheses and operational

definitions. This test was also meant to be used for elementary school children. In the same year, another researcher, Ludeman developed a science processes test, also aimed at elementary grade levels (Dillashaw and Okey, 1980).

The main shortcomings of the above stated tests were that most of them were based on specific curricula and evaluated a complex combination of skills rather than specific skills (Onwu and Mozube, 1992). Besides, the tests were said to have had uncertain validity because of the lack of external criteria by which to judge them (Molitor and George, 1976). In an attempt to separate the science process skills from a specific curriculum, Molitor and George (1976) developed a test of science process skills (TSPS), which focused on the inquiry skills of inference and verification, for grades four to six learners. This test was presented in the form of demonstrations. It was considered to be valid, but had a low reliability, especially for the inference subset, which had a reliability of 0.66 (Molitor and George, 1976). Most of the reviewed tests at the elementary level tended to deal with the basic science process skills only. None of them specifically addressed the assessment of higher order thinking skills. The review of these tests was helpful in selecting the methodology and format for the present study.

2.3.2 TEST DEVELOPMENT FOR SECONDARY SCHOOL LEVEL

At secondary school level, Woodburn, *et al* (1967) were among the pioneers of the development of science process skills tests for secondary school students (Dillashaw and Okey, 1980). They developed a test to assess secondary school learners' competence in methods and procedures of science. Tannenbaum (1971) developed a test of science processes, for use at middle and secondary school levels (grades seven, eight and nine). This test assessed skills of observing, comparing, classifying, quantifying, measuring, experimenting, predicting and inferring. It consisted of 96 multiple-choice questions. A weakness in this test related to the determination of criterion related validity, using a small sample of only 35 subjects. In addition, the scores obtained were compared to a rating scale prepared by the students' teacher, regarding competence in science processes skills (Lavinghousez, 1973), which could have been less accurate. The test was however

established and accepted by the educational community since it was unique and provided a complete testing manual (Lavinghousez, 1973).

Some flaws and weaknesses, in either the content, or the methodology used in the development of these early tests for secondary school level were identified. For instance, Dillashaw and Okey (1980) pointed out that in these early studies, attempts to measure knowledge of problem solving or the methods of science appear to combine tests of specific skills and scientific practices. Onwu and Mozube (1992) confirmed this observation by stating that most of the tests were curriculum oriented, and evaluated a complex combination of skills rather than specific skills. Like those developed for primary level, some of the tests were also said to have uncertain validity, because they did not have external criteria or a set of objectives by which to judge them (Molitor and George, 1976). Research evidence shows that, of the science curriculum projects for secondary schools, only the Biological Science Curriculum Study (BSCS) had a test specifically designed to measure process skills competence (Dillashaw and Okey, 1980). This test, referred to as the Biology Readiness Scale (BRS), was intended to provide a valid and reliable instrument to assess inquiry skills for improved ability grouping in the Biological Sciences Curriculum Study. The test, however, showed an exclusive use of Biological concepts and examples (Dillashaw and Okey, 1980).

Given some of the limitations as mentioned above, the researchers of the 80s and 90s developed further tests of integrated science process skills, which attempted to address some of the identified weaknesses. One of such tests was the Test of Integrated Science Processes (TISP), developed by Tobin and Capie (1982). This test was designed to examine grades six through college students' performance, in areas of planning and conducting investigations. The test items were based on twelve objectives, and it proved to have the ability to differentiate student abilities in inquiry skills. Padilla and Mckenzie (1986) developed and validated the test of graphing skills in science. The test was adjudged valid and reliable, but it only dealt with the process skills of graphing. Dillashaw and Okey (1980) however developed the more comprehensive Test of Integrated science Process Skills (TIPS), which included most of the integrated science process skills, such as identifying and controlling variables, stating hypotheses, designing

experiments, graphing and interpreting data, and operational definitions. The test was meant for use with middle grade and secondary school students. The test had a high reliability (0.89), and was also non-curriculum specific.

As a follow up on the TIPS, Burns, Okey and Wise (1985) developed a similar test, referred to as the Test of Integrated science Process Skills II (TIPS II). The test was based on the objectives and format of the original TIPS and it also had the same number of items (36). TIPS and TIPS II are usually used as equivalent subtests for pre and post assessment. Onwu and Mozube (1992) in the Nigerian setting also developed and validated a science process skills test for secondary science students. This test was also based on the format and objectives of the TIPS, developed by Dillashaw and Okey (1980). It was a valid test, with a high reliability (0.84).

Of all the tests stated above, only the science process skills test for secondary science students, developed by Onwu and Mozube (1992) was developed and validated in Africa. The few studies available in Africa show that researchers have been more interested in finding out the level of acquisition of some science process skills or in identifying the process skills inherent in a particular curriculum material (Onwu and Mozube, 1992).

Further more, none of the studies had so far attempted to determine test bias against possible sources such as the race, gender, school type, and location of the learners who may need to use their test. In this study, this aspect of sources of bias was taken into account during the development and validation of the test instrument.

2.4 SCIENCE PROCESS SKILLS TEST DEVELOPMENT IN SOUTH AFRICA.

A search of available tests of science process skills in South Africa, showed the need to develop such a test. Very little work has been done in the area of test development and validation, especially on the assessment of science process skills in schools. So far, there

is no published test of science process skills developed and validated in South Africa. This is in spite of the current reforms in the South African science education system, which is characterized by moves to promote science process skills acquisition, and inquiry-based investigative activities in science classes (Department of Education, 2002).

The Third International Mathematics and Science Study report by Howie (2001), indicated that the erstwhile South African science curriculum had only a minor emphasis on the explanation model and application of science concepts to solve problems. The report further indicated that the designing and conducting of scientific experiments and communicating scientific procedures and explanations are competencies hardly emphasized in science classes. Given the emphasis placed on the development of process skills in the new curriculum 2005, it became imperative to develop and validate a test instrument that would help assess learners' acquisition of those skills as a diagnostic measure, as well as a competence one.

2.5 CRITERIA FOR TEST DEVELOPMENT AND VALIDATION.

A major consideration in developing science process skills test is that of format (Dillashaw and Okey, 1980). Dillashaw and Okey pointed out that, while one requires students to demonstrate competence in science process skills, the problem of using hands-on procedures to assess skills acquisition could be a burdensome task. This is true in the context of large under-resourced classes. The paper and pencil group-testing format is therefore more convenient when assessing science process skills competence in large under-resourced science classes (Onwu and Mozube, 1992; Dillashaw and Okey, 1980), with the understanding that integrated science process skills are relatable to higher order thinking skills.

The general trend in the development of paper and pencil tests has been; the definition of the constructs and content to be measured, identification of the target population, item collection and preparation, pilot study, item review, main study, and data analysis with

regard to test characteristics (Ritter, Boone and Rubba, 2001; Gall and Borg, 1996; Nitko, 1996; Novak, Herman and Gearhart, 1996; Onwu and Mozube, 1992; Dillashaw and Okey, 1980; and Womer, 1968). A valid and reliable test should have test characteristics that fall within the accepted range of values, for each characteristic, such as; validity, reliability, discrimination index, index of difficulty, and readability, and it should not be biased against any designated sub-group of test takers. This section discusses the literature on test characteristics, and test bias.

2.5.1 TEST VALIDITY

Test validity, which is "the degree to which a test measures what it claims or purports to be measuring" (Brown, 1996, pp. 231), is a very important aspect of test construction. Validity was traditionally subdivided into; content validity, construct validity and criterion related validity (Brown, 2000; Wolming, 1998). Content validity includes any validity strategies that focus on the content of the test. To determine content validity, test developers investigate the degree to which a test (or item) is a representative sample of the content of the objectives or specifications the test was originally designed to measure (Brown 2000; Nitko, 1996; Wolming, 1998).

To investigate the degree of match, test developers enlist well-trained colleagues to make a judgment about the degree to which the test items matched the test objectives or specifications. This method was used in this study, to determine the content validity of the developed instrument. Criterion related validity involves the correlation of a test with some well respected outside measures of the same objectives and specifications (Brown 2000; Nitko, 1996). The Pearson product-moment is usually used for the correlation of scores. In this study, the TIPS (Dillashaw and Okey, 1980) was used to determine the criterion related validity of the developed test. The construct validity of a test involves the experimental demonstration that a test is measuring the construct it claims to be measuring. This may be done either through the comparison of the performance of two groups on the test, where one group is known to have the construct

under question and the other does not, or through the use of the "test, intervention, re-test" method (Brown, 2000; Wolming, 1998). Construct validity was determined in this study by comparing the performance of the different grade levels on the developed test, assuming that the learners in higher grades were more competent in science process skills (had the construct being measured) than those in lower grades.

Different researchers have different views on the acceptable test validity coefficient. For example, Adkins (1974), stated that the appropriateness of validity coefficients depends on several factors, and that "Coefficients of unit or close to unit, ordinarily are not attainable or expected", (Adkins, 1974; pp 33). She reiterated that the judgment of the value of validity coefficient is affected by the alternatives available. For instance, if some already existing test has a higher value than the new test, then the validity coefficient of the new test will be low compared to the existing test. The value of the validity coefficient also varies when the test is used for different purposes, and with varying characters of the subjects to which the test is given. She concluded that an important consideration is therefore to estimate validity for a group as similar as possible to the subjects for which the test is intended. Gall and Borg (1996), and Hinkle (1998) suggested a validity coefficient of 0.7 and more, as suitable for standard tests. Therefore, validity coefficients of 0.7 and more were considered to be appropriate for this study.

Other factors that may affect the validity of a test include its discrimination power, the difficulty level, its reliability, and the different forms of bias (Nitko, 1996). These factors were determined during the development of the test in this study, and are discussed in the following sections.

2.5.2 TEST RELIABILITY

Fundamental to the evaluation of any test instrument is the degree to which test scores are free from measurement error, and are consistent from one occasion to another, when the test is used with the target group (Rudner, 1994). Rudner stated that a test should be sufficiently reliable to permit stable estimates of the ability levels of individuals in the

target group. The methods used to measure reliability include; inter-rater reliability, test re-test method, alternate form (comparable form) reliability, and the internal consistency (split half) method. Of these methods, the internal consistency method is the most commonly used in test development research. The reason for its popularity is that, it accounts for error due to content sampling, which is usually the largest component of measurement error (Rudner, 1994). The test re-test is another method that is widely used by researchers to determine the reliability of a test. The disadvantage of using this method is that, examinees usually adapt to the test and thus tend to score higher in later tests (Adkins, 1974). Adkins advised that the test re-test method should be used as a last resort. The alternative form reliability is usually recommended by researchers. The problem with this method lies with the difficulty involved in finding equivalent tests for a specific assessment.

In this study, the internal consistency reliability was determined, because it was considered to be the most relevant and accurate method for the study. The alternative form reliability was also determined in this study, but it was primarily used for comparing the performance of learners on the developed test and a standard test, which was developed and validated in a different environment.

The recommended range of values for test reliability is from 0.7 to 1.0 (Adkins, 1974; Hinkle, 1998). Gall and Borg (1996) however proposed a reliability coefficient of 0.8 or higher to be sufficiently reliable for most research purposes. The latter coefficient was adopted in this study.

2.5.2.1. ESTIMATION OF STANDARD ERROR OF MEASUREMENT

The reliability of a test instrument can also be expressed in terms of the standard error of measurement (Gay, 1987). Gay contends that no procedure can assess learners with perfect consistency. It is therefore useful to take into account the likely size of the error of measurement involved in an assessment (Nitko, 1996). The standard error of measurement helps us to understand that the scores obtained on educational measures are

only estimates, and may be considerably different from an individual's presumed true scores (Gall and Borg, 1996). The standard error of measurement measures the distance of learners' obtained scores from their true scores (Nitko, 1996). A small standard error of measurement indicates a high reliability, while a large standard error of measurement indicates low reliability (Gay. 1987). In this study, the standard error of measurement was determined to further estimate the reliability of the developed instrument.

2.5.3 ITEM ANALYSIS

Item analysis is a crucial aspect of test construction, as it helps determine the items that need improvement or deletion from a test instrument. Item analysis refers to the process of collecting, summarizing, and using information from learners' responses, to make decisions about each assessment task (Nitko, 1996). One of the purposes of item analysis is to obtain objective data that signals the need for revising the items, so as to select and cull items from a pool (Nitko, 1996). This was the primary reason for doing item analysis in this study. The two central concepts in item analysis, especially in the context of this study are; index of difficulty and discrimination index, and they are discussed below.

2.5.3.1 DISCRIMINATION INDEX

Discrimination index of a test item describes the extent to which a given item distinguishes between those who did well in the test and those who performed poorly (Nitko, 1996). Discrimination index is determined by the difference between the proportion of high scorers who selected the correct option and that of low scorers who selected the correct option. Researchers contend that item discrimination indices of 0.3 and above are good enough for an item to be included in an assessment instrument (Adkins, 1974; Hinkle, 1998; Nitko, 1996).

Item discrimination index could also be based on the correlation between each item in a test and the total test score (Womer, 1968). This is referred to as the point bi-serial

correlation (the RPBI statistic). The larger the item-test correlation, the more an individual item has in common with the attribute being measured by the test (Womer, 1968). The use of the point bi-serial correlation indicates the direction and strength of the relationship between an item response, and the total test score within the group being tested. The RPBI statistic is recommended by many researchers as an effective way of selecting suitable items for a test, since it measures the discrimination power of the item in relation to that of the whole test. Womer suggested that item-test correlation indices of 0.4 and above indicate a relationship that is significant, and that such items should be retained in the final test. He however, recommended the inclusion of items with discrimination indices of as low as 0.2. The RPBI statistic was not considered in this study due to logistical reasons.

2.5.3.2 INDEX OF DIFFICULTY

Index of difficulty (difficulty level) refers to the percentage of students taking the test who answered the item correctly (Nitko, 1996). The larger the percentage of students answering a given item correctly, the higher the index of difficulty, hence the easier the item and vice versa.

Index of difficulty can also be determined by referring to the performance of the high scorers and the low scorers on a test (Croker and Algina, 1986). The former approach was adopted in this study.

Literature shows that the desired index of difficulty is around 50% (0.5) or within the range of 40 to 60% [0.4 – 0.6] (Nitko, 1996). It is recommended that items with indices of difficulty of less than 20% (0.20) and more than 80% (0.8) should be rejected or modified, as they are too difficult and too easy respectively (Nitko, 1996). Adkins (1974) suggested that a difficulty level should be about half way between the lowest and the highest scores. This suggestion agrees with that of Womer (1968), who proposed a difficulty level of 50% (0.5) to 55% (0.55) as being appropriate for the inclusion of a test

item into a test instrument. In this study, indices of difficulty within the range of 0.4 – 0.6 were considered appropriate for the developed test items.

2.5.4 TEST BIAS

The South African Educational System is characterized by a diversity of educational groupings and backgrounds that are likely to affect learners' academic performance. Language, gender, school types, race, and location of learners are among the educational groupings prevalent in South Africa. A test developed for such a diverse population of learners should seek to be relatively unbiased towards any of the different groups of the test takers.

Howe (1995), described bias as a kind of invalidity that arises relative to groups. A test is biased against a particular group if it disadvantages the group in relation to another (Howe, 1995, Childs, 1990). Hambleton and Rodgers (1995), defined bias as the presence of some characteristics of an item that result in differential performance for individuals of same ability, but from different ethnic, sex, cultural or religious groups. The most intuitive definition of bias is the observation of a mean performance difference between groups (Berk, 1982).

However, it should be noted that people differ in many ways. Finding a mean performance difference between groups does not necessarily mean that the test used is biased. The mean difference could either demonstrate bias or it could reflect a real difference between the groups, which could have resulted from a variety of factors, such as inadequate teaching and learning, or lack of resources. Nonetheless, in this study, mean performance differences between groups will be used to determine test bias.

While it is clear that a good test should not be biased against any group of test takers, literature shows that it is not easy to quantify test bias. Zieky (2002) contends that there is no statistic that one can use to prove that the items in a test or the test as a whole, is fair.

However, one way to assure test fairness according to Zieky (2002) is to build fairness into the development, administration, and scoring processes of the test.
This study therefore attempted to build in test fairness, during the test development process, to accommodate the diversity of learners prevalent in the South African education system.

2.5.4.1 Culture test bias

Intelligence is a distinctive feature of the human race, however, its manifestation and expression are strongly influenced by culture as well as the nature of the assessment situation (Van de Vijver and Hambleton, 1996). Any assessment is constructed and validated within a given culture (Van de Vijver and Poortinga, 1992). Assessments therefore contain numerous cultural references. The validity of an assessment tool becomes questionable when people from cultures that are different from the culture where the instrument was developed and validated use it. Brescia and Fortune (1988) pointed out in their article entitled "Standardized testing of American-Indian students" that, testing students from backgrounds different from the culture in which the test was developed magnifies the probability of invalid results, due to lack of compatibility of languages, differences in experiential backgrounds, and differences in affective dispositions toward handling testing environments between the students being tested and those for whom the test was developed and validated.

Pollitt *et al*, (2000) further pointed out that if context is not familiar, comprehension and task solutions are prevented, because culture, language and context may interact in subtle ways such that the apparently easy questions become impossible for the culturally disadvantaged students. Familiarity with the context is likely to elicit higher order thinking in solving a problem (Onwu, 2002)

What the literature suggests is that results from foreign developed performance tests may sometimes be considered unreliable and in turn invalid when used in a non discriminatory

way to test local learners (Adkins, 1974, Brescia and Fortune, 1988 and Pollitt and Ahmed, 2001). Such tests could therefore be considered to be culture and language biased against local learners. While it is true that culture free tests do not exist, culture fair tests are possible in the use of locally developed tests (Van de Vijver and Poortinga, 1992).

2.5.4.2 Gender test bias

Issues of gender bias in testing are concerned with differences in opportunities for boys and girls. Historically, females were educationally disadvantaged in South Africa, with the current political dispensation, there is concerted effort to attempt to narrow or eliminate the gender gap in the education system, by taking into account gender differences in the presentation of knowledge discussions. The development of gender sensitive tests is therefore likely to assist in this regard. In this study, an attempt was made to try to guard against gender test bias.

A test is gender biased if boys and girls of the same ability levels tend to obtain different scores (Childs, 1990). Gender bias in testing may result from different factors, such as the condition under which the test is being administered, the wording of the individual items, and the students' attitude towards the test (Childs 1990). Of these factors, the wording of the individual items is the one that is closely linked with test development.
Gender biased test items are items that contain; materials and references that may be offensive to members of one gender, references to objects and ideas which are likely to be more familiar to one gender, unequal representation of men and women as actors in test items, or the representation of one gender in stereotyped roles only (Childs, 1990). If test items are biased against one gender, the members of the gender may find the test to be more difficult than the other gender, resulting in the discrimination of the affected gender.

Gender bias in testing may also result from systemic errors, which involves factors that cannot be changed. For instance, Rosser (1989), found that females perform better on

questions about relationships, aesthetics and humanities, while their male counterparts did better on questions about sport, physical sciences and business. A joint study by the Educational Testing Services and the College Board (Fair Test Examiner, 1997), concluded that the multiple-choice format is biased against females, because females tend to be more inclined to considering each of the different options, and re-checking their answers than males. The study examined a variety of question types on advanced placement tests such as the Standard Assessment Test (SAT). They found that gender gap narrowed or disappeared on all types of questions except the multiple-choice questions. Test speediness has also been cited as one of the factors that bias tests against women. Research evidence shows that women tend to be slower than men when answering test questions (Fair Test Examiner, 1997). However, in this study, speed was not a factor under consideration.

2.5.4.3 Language test bias

An item may be language biased if it uses terms that are not commonly used nation wide, or if it uses terms that have different connotations in different parts of the nation (Hambleton and Rodger, 1995). Basterra (1999) indicated that, if a student is not proficient in the language of the test he/she is presented with, his/her test scores will likely underestimate his/her knowledge of the subject being tested.

Pollitt and Ahmed (2001) in their study on students' performance on TIMSS demonstrated that terms used in test questions are of critical importance to the learners' academic success. They pointed out that most errors that arise during assessment are likely to originate from misinterpretations when reading texts. Pollitt and Ahmed (2001) further explained that local learners writing tests written in foreign languages have to struggle with the problem of trying to understand the terms used, before they can attempt to demonstrate their competence in the required skill, and that, if the misunderstood term is not resolved, the learner may fail to demonstrate his or her competence in the required skill. They concluded that terms used in test questions are of critical importance to the learners' academic success.

In another study, Pollitt, Marriott and Ahmed (2000) interrogated the effect of language, contextual and cultural constraints on examination performance. One of the conclusions that they came up with, is that, the use of English words with special meaning can cause problems for learners who use English as a second language. The importance of language in academic achievement is supported by other researchers such as Kamper, Mahlobo and Lemmer (2003), who concluded that language has a profound effect on learners' academic achievements.

South Africa being a multi racial nation, is characterised by a diversity of languages of which eleven are considered as official languages that could be used in schools and other official places. Due to this diversity in languages, most learners in South Africa use English as a second or third language. As a result, they tend to be less proficient in English than the first English language users. It must however be understood that since the language of instruction in science classes in most schools in South Africa is either English or Afrikaans, it is assumed that the learners have some level of proficiency in these two languages. In light of the above literature, it was deemed necessary in this study to build in language fairness during the development of the test items. In order to estimate the language fairness of the developed test, it was necessary to determine its readability level. In consequence, the following passages discuss test readability.

2.5.5 TEST READABILITY

Readability formulae are usually based on one semantic factor [the difficulty of words], and one syntactic factor [the difficulty of sentences] (Klare, 1976). When determining the readability level of a test, words are either measured against a frequency list, or are measured according to their length in characters or syllables, while sentences are measured according to the average length in characters or words (Klare, 1976).

Of the many readability formulae available, the Flesch reading ease scale (Klare, 1976) is the most frequently used in scientific studies, due to the following reasons; first, it is

easy to use, since it does not employ a word list, as such a list may not be appropriate for science terminologies. Second, it utilizes measurement of sentence length and syllable count, which can easily be applied to test items. Lastly, the Flesch measure of sentence complexity is a reliable measure of abstraction (Klare, 1976). The latter reason is very important because the comprehension of abstract concepts is a major problem associated with science education. The formula also makes adjustments for the higher end of the scale (Klare, 1976). The Flesch scale measures reading from 100 (for very easy to read), to 0 (for very difficult to read). Flesch identified a '65' score as the plain English score (Klare, 1976). In this study, the Flesch reading ease formula was therefore selected for the determination of the readability level of the developed test instrument.

Despite the importance attached to readability tests, critics have pointed out several weaknesses associated with their use. In recent years, researchers have pointed out that readability tests can only measure the surface characteristics of texts. Qualitative factors like vocabulary difficulty, composition, sentence structure, concreteness and abstractness, and obscurity and incoherence cannot be measured mathematically (Stephens, 2000). Stephens (2002) also indicated that materials which receive low grade-level scores, might be incomprehensible to the target audience. He further argued that because readability formulae are based on measuring words and sentences, they cannot take into account the variety of resources available to different readers, such as word recognition skills, interest in the subject, and prior knowledge of the topic.

Stephens (2000) contends that the formulae does not take into account the circumstances in which the reader will be using the text, for instance, it does not measure psychological and physical situations, or the needs of people for whom the text is written in a second or additional language. He suggested that a population that meets the same criteria for first language must be used to accurately assess the readability of material written in a second or additional language.

In this study test readability level was determined to provide an estimation of the degree to which the learners would understand the text of the developed instrument, so that

learners may not find the test to be too difficult due to language constraints. A reading level of 60 – 70 was considered to be easy enough for the learners to understand the text of the test instrument. However, it was preferable for the readability level of the developed test instrument to be on the higher end of the readability scale (≤ 70) due to the reasons advanced above.

CHAPTER 3
RESEARCH METHODOLOGY

This chapter discusses the research design, population and sample description, instrumentation, the pilot study, the main study, statistical procedures for data analysis, and ethical issues.

3.1 RESEARCH DESIGN

The research was an ex post facto research design, involving a test development and validation study that used a quantitative survey type research methodology. This research design was found to be suitable for this kind of study.

3.2 POPULATION AND SAMPLE DESCRIPTION

The population of the study included all FET learners in the Limpopo province of South Africa. The sample used in the study was derived from the stated population. Specifically, the sample comprised 1043 science learners in high schools in the Capricorn district of the Limpopo province. The pilot study involved 274 subjects, selected from two rural and two urban schools that were sampled from two lists of rural and urban schools found in the Capricorn district of the Limpopo province. The main study involved 769 subjects selected from six schools sampled from the above-mentioned lists of schools. The selected sample consisted of grade 9, 10, and 11 science learners from different school types, gender, race, and location in the respective schools. The involvement of different groups of learners was necessary for the comparison of the test results, so as to determine the sensitivity of the test instrument. The schools that participated in the pilot study were not involved in the main study.

The following method was used to select the schools that participated in the study. Two lists consisting of the urban and rural schools in the Capricorn district were compiled. Two schools were randomly selected from each list, for use in the Pilot study. The schools that participated in the two trials of the pilot study are shown on table 3.1 below. PR was the code for the rural schools that were used in the pilot study, while PU represented the urban schools used. The table also shows the school type and the race of the learners.

TABLE 3.1 SCHOOLS THAT PARTICIPATED IN THE PILOT STUDY

School Code	School	Location	School type	Race
PR1	High school 1	Urban	Model C[1]	Black
PR2	High school 2	Rural	DET[2]	Black
PU1	High school 3	Rural	DET	Black
PU2	High school 4	Urban	Model C	Black

For the main study, two lists comprising formerly Model C and Private schools were drawn from the remaining list of urban schools. The list of rural schools comprised formerly DET schools only. The division of the urban schools into the stated school types led the formation of three school lists, consisting of formerly model C schools, private schools, and DET schools (all from rural schools).

The schools that participated in the main study were selected from these three lists as follows; First, schools with white learners only were identified from the list of formerly model C schools, as there were no such schools on the other two lists, and two schools were randomly selected from the identified schools. Second, schools comprising white and black learners were identified from the list of formerly model C schools, for the same reason as given above. Two schools were randomly selected from the identified schools.

Footnote:

1. *Model C schools are schools which were previously advantaged under the apartheid regime*
2. *DET schools are schools which were previously managed by the Department of Education and Training, and were disadvantaged under the apartheid regime.*

Third, two schools with black learners only were randomly selected from the remaining list of formerly model C schools.

Lastly, two schools were randomly selected from each of the lists of private and rural schools. All the learners from the private and rural schools selected were black. In total, ten schools comprising two formerly model C schools with white learner, two formerly model C schools with mixed learners, two formerly model C schools with black learners, two private schools with black learners, and two formerly DET rural schools with black learners were selected for use in the main study.

The two formerly model C schools with white learners only withdrew from the study as the learners could not write an English test, since they were Afrikaans Speaking learners. The researcher was requested to translate the developed test into Afrikaans, but was unable to do so during the study period. One private school also withdrew because the principal did not approve of the study, and one formerly model C school with black learners could not participate in the study at the time of test administration, since the school had just lost a learner, and preparations for the funeral were under way. Finally, only six schools were able to participate in the main study, and their names, school type, and races of learners are indicated on table 3.2 below. The ratio of white to black learners in the racially mixed schools was approximately 50:50 and 70:30 respectively.

TABLE 3.2 SCHOOLS THAT PARTICIPATED IN THE MAIN STUDY

School Code	School	Location	School type	Race
A	High school 5	Urban	Model C	White:Black/50:50
B	High school 6	Urban	Model C	Black
C	High school 7	Urban	Private	Black
D	High school 8	Rural	DET	Black
E	High school 9	Rural	DET	Black
F	High school 10	Urban	Model C	White:Black/70:30

3.3 INSTRUMENTATION

The instrument used in the study was a test of integrated science process skills, developed by the researcher. The instrument was used to collect data that was used for the determination of its test characteristics, and for the comparison of the performance of different groups of learners on the test. The Test of Integrated Science Process Skills (TIPS), developed by Dillashaw and Okey (1980), was also used for the determination of the concurrent validity and the alternative form reliability of the developed instrument.

3.3.1 PROCEDURE FOR THE DEVELOPMENT AND WRITING OF THE TEST ITEMS.

The South African science curriculum statements for the senior phase of the GET, and the FET bands, as well as prescribed textbooks and some teaching material were reviewed and analysed, to ascertain the inclusion of the targeted science process skills, and the objectives on which the test items were based.

A large number of test items was initially constructed from various sources, such as locally prepared past examinations and tests, science selection tests, standard achievement tests, textbooks, and from day to day experiences. The items were referenced to a specific set of objectives (Onwu and Mozube, 1992; Dillashaw and Okey, 1980). These objectives are related to the integrated science process skills of; identifying and controlling variables, stating hypotheses, making operational definitions, graphing and interpreting data, and designing investigations. The stated integrated science process skills are associated with planning of investigations, and analysis of results from investigations. The objectives to which the test items were referenced are shown on table 3.3 below.

TABLE 3.3. OBJECTIVES UPON WHICH TEST ITEMS WERE BASED

Science process skill measured	**Objective**
Identifying and controlling variables	1. Given a description of an investigation, identify the dependent, independent and controlled variables.
Operational definitions	2. Given a description of an investigation, identify how the variables are operationally defined.
Identifying and controlling variables	3. Given a problem with a dependent variable specified, identify the variables, which may affect it.
Stating hypotheses	4. Given a problem with dependent variables and a list of possible independent variables, identify a testable hypothesis.
Operational definitions	5. Given a verbally described variable, select a suitable operational definition for it.
Stating hypotheses	6. Given a problem with a dependent variable specified. Identify a testable hypothesis.
Designing investigations	7. Given a hypothesis, select a suitable design for an investigation to test it.
Graphing and interpreting data	8. Given a description of an investigation and obtained results/data, identify a graph that represents the data.
Graphing and interpreting data	9. Given a graph or table of data from an investigation, identify the relationship between the variables.

The above objectives were adopted from the 'Test of integrated Science Process Skills for Secondary schools' developed by F.G Dillashaw and J. R. Okey (1980), and also used in the Nigerian context by Onwu and Mozube (1992), with a slight modification to objective 1.

The items comprising the test instrument were designed in such a way that tried to assure that they do not favour any particular science discipline, gender, location, school type, or race. In order to avoid the use of items that are content specific, each test item was given to two science educators at the university of Limpopo, as judges, to determine whether the item was content specific to any particular science discipline or not, before it was included in the draft test instrument.

Furthermore, in attempting to minimize test bias against gender, race, location, and school type, the same science educators were asked to judge whether:

(i) the references used in the items were offensive, demeaning or emotionally charged to members of some groups of learners
(ii) reference to objects and ideas that were used were likely to be more familiar to some groups of learners than others
(iii) some groups of learners were more represented as actors in test items than others, or
(iv) certain groups of learners were represented in stereotyped roles only

Initially, about 8 to 9 items, referenced to each of the stated objectives (Table 3.3) were selected in this manner. The total number of items selected were 76 multiple-choice test items, each having four optional responses. Only one of the four optional responses was correct. Care was taken to assure that the distracters were incorrect but plausible. These items formed the first draft instrument. The number of selected test items reduced as the instrument went through the various development stages. The format of the test instrument was modelled after the test of integrated science process skills (TIPS) developed by Dillashaw and Okey (1980).

TABLE 3.4. LIST OF INTEGRATED SCIENCE PROCESS SKILLS MEASURED, WITH CORRESPONDING OBJECTIVES AND NUMBER OF ITEMS SELECTED

	INTEGRATED SCIENCE PROCESS SKILL MEASURED	**OBJECTIVES**	**NUMBER OF ITEMS**
A	Identifying and controlling variables	1 and 3	17
B	Stating hypotheses	4 and 6	17
C	Operational definitions	2 and 5	17
D	Graphing and interpreting data	8 and 9	17
E	Designing investigations	7	8
Total	**5**	**9**	**76**

3.3.2 FIRST VALIDATION OF THE TEST INSTRUMENT

The first draft test instrument was tested for content validity by six peer evaluators (raters) who comprised two Biology lecturers, two Physics lecturers, and two Chemistry lecturers from the University of Limpopo. These raters were given the test items and a list of the test objectives, to check the content validity of the test by matching the items with the corresponding objectives. The content validity of the instrument was obtained by determining the extent to which the raters agreed with the test developer on the assignment of the test items to the respective objectives (Dillashaw and Okey, 1980; Nitko. 1996). From a total of 456 responses (6 raters X 76 items used), 68 percent of the rater responses agreed with the test developer on the assignment of the test items to objectives. This value was pretty low for content validity. It should however be noted that this validation of the instrument was done prior to the administration of the pilot study. Therefore, this value changed after the item reviews and modifications that resulted from the pilot study item analysis.

The raters were also asked to provide answers to the test items so as to verify the accuracy and objectivity of the scoring key. The analysis of their responses showed that 95 percent of the raters' responses agreed with the test developer on the accuracy and objectivity of the test items. The items on which the raters did not select the same answers as the test developer were either modified or discarded. Further more, an English lecturer was asked to check the language of the test items, in terms of item faults, grammatical errors, spelling mistakes and sentence length. The instrument was also given to some learners from grades nine (9), ten (10), and eleven (11) to identify difficult or confusing terms or phrases from the test items. The recommendations from the lecturer and the learners were used to improve the readability of the test instrument.

All the comments from the different raters were used to revise the test items accordingly. Items that were found to have serious flaws, especially the ones where the raters did not agree with the test developer on assigning them to objectives, were discarded. This first validation of the items led to the removal of several unsuitable items.

By the end of the review process, 58 items were selected, and they constituted the second draft of the test instrument, which was administered to learners in the pilot study.

3.4 PILOT STUDY

The developed instrument was initially administered to learners in a pilot study, which consisted of two phases (trials). These phases were referred to as the first trial and the second trial studies, as discussed below.

3.4.1 PURPOSE OF THE PILOT STUDY

The purpose of the first trial study was first, to establish the duration required by the learners to complete the test. The duration for the test was not specified during the administration of the test in the first trial study. Instead, a range of time in which the learners completed the test was determined. The first learner to complete the test took 30 minutes, while the last one took 80 minutes. It was therefore established that for the 58 item test used in the pilot study, the learners required more than two school periods (of about 70 minutes) to complete the test. Secondly, the data collected from the first trial study was used to find out whether there were any serious problems with the administration of the test instrument and management of the results.

The purpose of the second trial study was to try out the test instrument on a smaller sample, so as to determine its test characteristics. These test characteristics included the reliability, discrimination index, index of difficulty, the readability level, and the item response pattern of the developed instrument. Most importantly, the data from the second trial study, and the test characteristics obtained were used to cull the poor items from the pool of test items selected..

3.4.2 SUBJECTS USED IN THE PILOT STUDY.

The subjects used in the pilot study comprised a total of 274 learners in grades 10, and 11, from four selected schools in the Capricorn district. The first trial study involved 124 science learners from one rural and one urban school, while the second trial study used 150 science learners, also from one urban and one rural school. The participating classes were randomly selected from grade 10 and 11 learners in each school. It was not necessary to involve the different categories of learners used in the main study, during the pilot study because performance comparisons were not required at this stage.

3.4.3 ADMINSTRATION OF THE PILOT STUDY

The researcher applied for permission to administer the test to learners from the provincial department of Education through the district circuit. After obtaining permission from the department, the researcher sought permission from the respective principals of the schools that were selected for the study. A timetable for administering the test to the various schools was drawn and agreed upon with the respective principals. Two research assistants were hired to assist with the administration of the test to learners. On the appropriate dates, the researcher and the assistants administered the developed test instrument to learners. Prior to each administration of the test, the purpose of the study and the role of the learners were thoroughly explained to the subjects. They were also informed of their right to decline from participating in the study if they so wished.

After the administration of the test in the four schools used in the pilot study, the scripts were scored by allocating a single mark for a correct response, and no mark for a wrong, omitted, or a choice of more than one response per item. The total correct score was determined, and the percentage of the score out of the total number of possible scores (the total number of items) was calculated. Both the raw scores and the percentages for each subject were entered into a computer for analysis. Codes were used to identify the subjects and the schools where they came from. The test characteristics of the instrument were determined as discussed below.

3.4.4 RESULTS AND DISCUSSIONS FROM THE PILOT STUDY.

3.4.4.1 ITEM RESPONSE PATTERN.

The results from the pilot study were analyzed, and an item response pattern was determined as shown on table 3.5 below. The table shows that several items were very easy. For instance, the data shows that almost all the participants selected the correct option for items 1, 10, and 29 and these items measured the skills of identifying and controlling variables. Such items were too easy and they were subsequently replaced. Some items had bad distracters, whereby nobody selected the particular option. Examples of such distracters included options C and D for item 5; options A and D for item 7 and many others (Table 3.5). All distracters, which were not selected by anyone, were either modified or replaced. On the other hand, very few participants selected the correct options for items 22, 27, 30, 31, 32, 33, 34, 43, 45, 50, and 56. These items tested skills of operational definitions and designing experiments. Such items were considered too difficult and were either modified or replaced.

TABLE 3.5 SUMMARY OF THE ITEM RESPONSE PATTERN FROM THE PILOT STUDY.

Q	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**
A	0	12	6	6	**126**	0	0	6	12	**150**	6	**132**	6	**90**	18	6	12	0	18	0	0	42	0	**138**	24	**138**	**72**	0	0
B	**150**	18	42	**102**	24	6	**90**	0	**102**	0	48	6	12	42	6	**108**	**138**	24	12	6	24	**72**	24	0	6	6	12	0	**144**
C	0	24	6	12	0	**120**	60	12	0	0	12	6	**132**	12	18	0	0	**120**	6	**138**	**120**	30	30	6	**114**	6	24	18	6
D	0	**96**	**96**	30	0	24	0	**132**	36	0	**84**	6	0	6	**108**	36	0	6	**114**	6	6	6	**96**	6	6	0	42	**132**	0
Q	**30**	**31**	**32**	**33**	**34**	**35**	**36**	**37**	**38**	**39**	**40**	**41**	**42**	**43**	**44**	**45**	**46**	**47**	**48**	**49**	**50**	**51**	**52**	**53**	**54**	**55**	**56**	**57**	**58**
A	37	18	26	24	**61**	0	0	27	31	**100**	26	**72**	58	20	10	36	**102**	6	**108**	0	0	22	0	28	14	**78**	43	0	66
B	**54**	10	**52**	30	34	34	65	0	**91**	0	**78**	29	**92**	32	20	48	48	**144**	0	36	49	37	**124**	0	36	36	**72**	0	6
C	35	46	33	**76**	38	**80**	**85**	31	0	0	32	26	0	**72**	**108**	0	0	0	42	**78**	**65**	**80**	0	36	**84**	36	35	**108**	**78**
D	24	**76**	39	20	17	30	0	**92**	28	50	14	23	0	26	12	**66**	0	0	0	36	36	11	26	**86**	16	0	0	42	0

KEY:

Bold = Correct option; Q = item number

A,B,C,D = optional responses for each test item

The total number of subjects N = 150

Items with bad distracters but were considered appropriate for the test, were isolated and administered without the options to a group of grade 10 learners from one of the urban schools used in the pilot study. These learners were asked to provide their own responses. The wrong responses that appeared frequently for each of the selected items were used to replace the inappropriate distracters.

3.4.4.2 DISCRIMINATION AND DIFFICULTY INDICES

The statistical procedures and formulae used in the main study and described in sections 3.7.3 and 3.7.4 were applied to determine the discrimination and difficulty indices of the test items used in the pilot study. Analysis of the difficulty indices of these items showed that, about 40% of the items had difficulty indices of more than 0.8, with an average index of difficulty of 0.72 (Table 3.6).

TABLE 3.6 DISCRIMINATION AND DIFFICULTY INDICES FROM THE PILOT STUDY RESULTS.

Item no.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
Discrim	0	0.3	0.5	0.5	0.4	0.3	0.7	0.1	0.3	0	0.3	0.2	0.3	0.7	0.3	0.2	0.1	-1	0.5
Diff.	1	0.7	0.8	0.7	0.7	0.7	0.6	0.9	0.5	1	0.5	0.8	0.8	0.3	0.7	0.8	0.9	0.9	0.6

Item no.	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38
Discrim	0.1	0.5	0.3	0.5	0.1	0.3	0.3	0.5	0.1	0.6	0.1	0.5	0.5	0.5	-1	0.5	0.3	0.6	0.1
Diff.	0.9	0.7	0.2	0.6	0.9	0.6	0.8	0.4	0.9	0.7	0.9	0.7	0.4	0.7	0.9	0.7	0.7	0.7	0.9

Item no.	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58
Discrim	0.5	0	0.1	0.2	0.3	0.7	0.3	0.3	0.3	0.5	0.3	0.1	0.5	0.3	0.5	0.1	0.3	0.3	0.3	0.1
Diff.	0.6	1	0.9	0.8	0.8	0.3	0.7	0.7	0.5	0.6	0.5	0.9	0.7	0.2	0.6	0.9	0.7	0.8	0.7	0.7

Key
Discrim = Discrimination index
Diff = Index of difficulty
Number of subjects = 150

Average discrimination index = 0.32

Average Index of difficult y =0 .722

A test instrument with an index of difficulty of more than 0.6 is considered to be too easy (Nitko, 1996). In this study, a difficulty index range of 0.4 to 0.6 was considered appropriate for the inclusion of an item in the test instrument. It was therefore necessary to modify, replace, or simply discard the items that were outside this range.

The discrimination index is a very important measure of the item quality, in identifying learners who posses the desired skills and those who do not. The results from the trial study showed that about 31% of the items had low discrimination indices [less than 0.3] (Table 3.6). This could have resulted from the large number of poor distracters observed in the test items. Items 18 and 34 had negative discrimination indices, which means that low scorers found them to be easier, while the high scorers found them to be difficult. These items were discarded. Items 1, 10 and 40 had discrimination indices of zero (0), which means that the items could not discriminate at all between learners who had the desired skills and those who did not. These items were also discarded. The rest of the items had good discrimination indices, and were therefore retained in the draft instrument. The overall discrimination index of the instrument was 0.32, which was within the acceptable range of values for this test characteristic (see Table 3.6). The removal of the items that did not discriminate well, improved the overall discriminating power of the instrument.

3.4.4.3 RELIABILITY AND READABILTY OF THE INSTRUMENT

The data were further analysed to determine the reliability of the instrument using the split half method of determining the internal consistency reliability, and it was found to be 0.73. While this value falls within the accepted range of values for this test characteristic, it was still on the lower end, meaning that the test was not very reliable. The readability of the instrument was determined using the Flesch reading ease formula, which was found to be 59. This readability level falls below the accepted range of values for this test characteristic, which suggests that the test was difficult to read.

The table below shows the summary of the test characteristics obtained from the pilot study results, and are compared with the accepted range of values as determined by literature.

TABLE 3.7 SUMMARY OF THE PILOT STUDY RESULTS

GRADE	VALUE	ACCEPTABLE RANGE OF VALUES
Content validity	0.68	≥0.70
Reliability	0.73	≥0.70
Discrimination index	0.32	≥0.3
Index of difficulty	0.72	0.4 – 0.6
Readability	59	60 - 70

Test characteristics values obtained from the pilot study mostly fell outside the acceptable range of values, as shown on table 3.7 above. They were therefore considered to be un-satisfactory. Items with poor test characteristics were either modified or discarded. At the end of the pilot study analysis and review, only 31 items were selected for use in the main study.

3.5 SECOND VALIDATION OF THE TEST INSTRUMENT

As stated earlier (section 3.3.2), the initial validation of the test instrument showed that the instrument had a low content validity (0.68). The reviews and modifications that followed the initial validation and the pilot study resulted in a relatively different pool of items. It was therefore necessary to determine the content validity the instrument again before it could be used in the main study.

The procedure for validating the instrument was carried out as described in the initial validation of the instrument (section 3.3.2), using the same raters. From a total of 186 responses (6 raters X 31 items used), 98 percent of the rater responses agreed with the test developer on the assignment of the test items to objectives, and 100 percent of the

raters' responses agreed with the test developer on the accuracy and objectivity of the test items. The determination of these values was done as follows:

Content validity

182 * 100 = 97.84946
186 (182 = No. of responses that agreed with the test developer)
(186 = Total No. of responses)

Objectivity of items

186 = 100%
186

This concurrence of raters was taken as evidence of content validity and objectivity of the scoring key.

3.6 MAIN STUDY

3.6.1 NATURE OF THE FINAL TEST INSTRUMENT

After carrying out the various reviews of the test items, a final instrument, which was a paper and pencil test consisting of 31 multiple-choice questions was developed. Each question carried four optional responses, where only one response was correct and the other three options served as distracters.

The multiple-choice format was perceived as the most appropriate format for this study despite some of the weakness associated with the format, such as no provision for the reasons for the selection of a particular option. But this essentially was not the intention of the study. The study was to develop a test of integrated science process skills, and to this end, the multiple-choice format can be used to compare performance from class to class and from year to year.

Multiple-choice questions (MCQs) are widely used in educational systems. For instance, Carneson, J. *et. al.* (2003) stated that, a number of departments at the University of Cape Town (UCT), have been using multiple-choice questions for many years and the experience has generally been that, the use of multiple-choice questions has not lowered the standards of certification, and that there is a good correlation between results obtained from such tests and more traditional forms of assessment, such as essays. Multiple-choice questions are credited with many advantages, which tend to offset their weakness, and the following are some of the advantages of using the multiple-choice format.

- Multiple-choice questions can be easily administered, marked and analysed using computers, especially for large classes. Web-based formative assessment can easily be done using multiple-choice questions, so that learners from different areas may access the test, and that they may get instant feedback on their understanding of the subject involved.
- The scoring of multiple-choice questions can be very accurate and objective, so variations in marking due to subjective factors are eliminated, and MCQs do not require an experienced tutor to mark them (Higgins and Tatham, 2003).
- Multiple-choice questions can be set at different cognitive levels. They are versatile if appropriately designed and used (Higgins and Tatham, 2003).
- Multiple-choice questions can provide a better coverage of content and assessment can be done in a short period of time.
- Multiple-choice questions can be designed with a diagnostic end in mind, or can be used to detect misconceptions, through the analysis of distracters.
- Multiple-choice questions can easily be analysed statistically, not only to determine the performance of the learners, but the suitability of the question and its ability to discriminate between learners of different competencies.
- In multiple-choice questions, the instructor "sets the agenda" and there are no opportunities for the learner to avoid complexities and concentrate on the superficial aspects of the topic, as is often encountered in Essay-type questions.

- Multiple-choice questions focus on the reading and thinking skills of the learner, and does not require the learner to have writing skills, which may hinder the demonstration of competence in the necessary skills.

The decision to use the multiple-choice format was influenced by the above stated advantages.

Table 3.8 below displays the item specification. It shows the number of questions allocated to each of the integrated science process skills considered in this study. The table shows that the skill of graphing and interpreting data had more items (9) than other skills. The reason for this was that the skill contains several other sub-skills, such as identifying relationships, reading graphs, drawing relevant graphs, describing data, etc, which needed to be taken into account, while other skills do not have so many sub-skills.

TABLE 3.8 ITEM SPECIFICATION TABLE

	Integrated Science Process Skill	**Objectives**	**Number of items**
A	Identifying and controlling variables	1 and 3	2, 6, 19, 25, 28, 29, 30 = 7
B	Stating hypotheses	4 and 6	8, 12, 16, 20, 23, 26 = 6
C	Operational definitions	2 and 5	1, 7, 10, 18, 21, 22 = 6
D	Graphing and interpretation of data	8 and 9	4, 5,9, 11,14, 17, 24, 2731 =9
E	Experimental design	7	3, 13, 15 = 3
	5 Integrated science process skills	**9 objectives**	Total number of items **= 31**

The items associated with each of the nine objectives are shown on Table 3.8 below. Each objective was allocated three (3) items, except for objectives 1 and 9 that had 4 and 5 items respectively. The reason for this discrepancy is the number of sub-skills subsumed under the skills measured by these objectives.

TABLE 3.9 ALLOCATION OF ITEMS TO THE DIFFERENT OBJECTIVES

Objective on which the item was based	**Number of items allocated to it.**
1. Given a description of an investigation, identify the dependent, independent, and controlled variables.	2, 28, 29, 30
2. Given a description of an investigation, identify how the variables are operationally defined.	7, 18, 21
3. Given a problem with a dependent variable specified, identify the variables that may affect it.	6, 19, 25
4. Given a problem with dependent variables and a list of possible independent variables, identify a testable hypothesis.	20, 23, 26
5. Given a verbally described variable, select a suitable operational definition for it.	1, 10, 22
6. Given a problem with a dependent variable specified, identify a testable hypothesis.	8, 12, 16
7. Given a hypothesis, select a suitable design for an investigation to test it.	3, 13, 15
8. Given a description of an investigation and obtained results/data, identify a graph that represents the data.	9, 14, 24
9. Given a graph or table of data from an investigation, identify the relationship between the variables.	4, 5, 11, 17, 27

3.6.2 SUBJECTS USED IN THE MAIN STUDY.

The final test instrument was administered to 769 learners in grades 9, 10, and 11, from the six selected schools, comprising formerly DET schools, formerly model C schools, and private schools coming from urban and rural areas, as shown on table 3.2. The subjects were black and white boys and girls. There were 264 grade 9 learners, 255 grade 10 learners, and 250 grade 11 learners.

3.6.3 ADMINISTRATION OF THE MAIN STUDY

The method used to administer the test in the pilot study was used in the main study (3.4.3). The instrument was administered to grade 9, 10, and 11 science learners in all the six selected schools. The duration of the test for every session was two school periods, and it was sufficient for all the subjects involved.

In each school, the principal, in collaboration with class teachers decided on the classes to which the test was to be administered, according to their availability. In other words, the school authorities identified the classes which had double periods, and did not have other serious school commitments, such as writing a test, performing a practical, going on a field trip etc, and released them for the administration of the developed test. One school was randomly selected from the six selected schools in which the developed test was administered concurrently with the TIPS instrument (Dillashaw and Okey, 1980), for the determination of the alternative form reliability and concurrent validity. Arrangements were made with the principal of the school to allow the learners to write the test in the afternoon, after the normal school schedule, to allow for the extra time that was required to complete both tests. Two research assistants were hired to help with the administration of the test, in all the selected schools.

3.6.4 MANAGEMENT OF THE DATA FROM THE MAIN STUDY

The test items were scored as described in section 3.4.3. Each school was given a letter code, while each learner was given a number code associated with the school code, according to the grade levels. The learner code therefore reflected the school, the grade and the learner's number. For instance, C1025, would mean learner number 25, in grade 10, at Capricorn High School. The total score and the percentage for each learner was fed into a computer, against the learner's code number. Six more research assistants were hired and trained to assist with the scoring, and capturing of the results into the computer. The entered scores were analysed statistically using the micro-soft excel, and SPSS for windows programs, as follows:

First, data from all the 769 subjects were analysed to determine the item response pattern, the discrimination index, and the index of difficulty of the items, and consequently those of the test instrument as a whole.

Second, data from 300 subjects, comprising 100 learners randomly selected from each grade level, were used to determine the internal consistency reliability of the test instrument. The Pearson product moment coefficient and the Spearman brown prophecy formulae were used for this computation. The standard error of measurement was also determined, using the same sample, to further estimate the reliability of the instrument (Gay, 1987).

Third, the performance of 90 learners (comprising 30 subjects randomly selected from each grade level), on both the developed test and the TIPS (Dillashaw and Okey, 1980), was correlated using the Pearson product moment coefficient, to determine the concurrent validity and the alternative form reliability of the instrument. This computation was also used to compare the performance of the learners on both tests, to confirm or nullify the claim that foreign developed tests sometimes posed challenges for local learners. The 90 learners used in this analysis were from the school where the developed test and the TIPS were concurrently administered.

Fourth, the readability level of the instrument was determined using the Flesch reading ease formula, while the grade reading level was determined using the Flesch-Kincaid formula (section 4.5).

Lastly, the performance of learners from the different school types (formerly model C, formerly DET, and private schools), gender (girls and boys), race (whites and blacks), location (rural and urban schools), and grades (grade 9, 10 and 11) was compared using tests of statistical significance [t-test for independent and dependent samples, and simple analysis of variance (ANOVA)]. This was done to determine whether the learners‘ performances were significantly differences at $p \leq 0.05$. This computation was used to determine whether the test instrument had significant location, race, school type, or gender bias. For each comparison, the same number of subjects was randomly selected from the respective groups, as explained in section 4.6. Provision was made on the question paper for learners to indicate demographic information required for the study.

3.7 STATISTICAL PROCEDURES USED TO ANALYSE THE MAIN STUDY RESULTS

3.7.1 MEAN AND STANDARD DEVIATION

The grade means and standard deviations for the different groups of learners were determined using the computer, and confirmed manually, by using the formulae given below.

Mean

$X = \frac{\sum x}{N}$ Where X = Mean score

$\sum x$ = Sum of the scores obtained

N = Total number of students who wrote the test

Standard deviation

SD = $\sqrt{s^2}$ Where: s = variance

SD =Standard deviation

3.7.2 ITEM RESPONSE PATTERN

The item response pattern shows the frequency of the choice of each alternative response in a multiple-choice test. To determine the item response pattern of the main study, the subjects were divided into high, middle, and low scorer performance categories. These performance categories were determined by first arranging all the subjects' scores on the test in a descending order. Secondly, the subjects whose scores fell in the upper 27% of the ranking were considered to be high (H) scorers, while those whose scores fell in the lower 27% of the ranking were considered to be low (L) scorers. The remaining subjects were considered to be medium scorers.

Each test item was assigned a number, and the number and percentage of learners who selected each option was determined, for each item. The number of learners who omitted the item or marked more than one option (error) for each item, was also shown, for each of the high, medium, low, and total score groups.

The learners' responses for each item were then analysed. If too many test takers selected the correct option to an item, then the item was too easy. Conversely, if too many selected the wrong options, then the item was too difficult. Such items were either reviewed or discarded. Similarly, if too many test takers, especially those in the high score group selected a distracter, then it was considered to be an alternative correct response, and was therefore modified or discarded. If very few or no test takers selected a distracter, then it was considered not plausible, and was discarded.

3.7.3 ITEM DISCRIMINATION INDEX

The discrimination index of each item was obtained by subtracting the proportion of low scorers who answered the question correctly, from the proportion of high scorers (section 3.7.2) who answered the question correctly (Trochium, 1999). A good discrimination item is one where a bigger proportion of the high scorers selected the correct option than the low scorers. The higher the discrimination index, the better the discriminability of the item. The following formula was used to determine the discrimination index of the items.

$$D = \frac{R_H}{n_H} - \frac{R_L}{n_L}$$

Where;
- D = item discrimination index.
- R_H = number of students from the high scoring group who answered the item correctly.
- R_L = number of students from the low scoring group who answered the item correctly.
- n_H = Total number of high scorers.
- n_L = Total number of low scorers.

3.7.4 INDEX OF DIFFICULTY

The index of difficulty was determined by calculating the proportion of subjects taking the test, who answered the item correctly (Nitko, 1996). To obtain the index of difficulty (p), the following formula was used;

$p = \frac{R*100}{n}$ Where; p = index of difficulty.
n = total number of students in the high scoring and low scoring groups.
R = number of high and low scoring students who answered the item correctly.

3.7.5 RELIABILITY OF THE INSTRUMENT

The reliability of the test instrument was determined in two ways: first, by using the split half method of determining the internal consistency of the test, where the test items were split into odd and even-numbered items. The odd-numbered items constituted one half test, and the even-numbered items constituted another half test, such that, each of the sampled students had two sets of scores. The scores obtained by each subject on the even-numbered items were compared and correlated with their scores on the odd-numbered items, using the Pearson product–moment coefficient (Mozube, 1987; Gay, 1987), as follows;

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

Where;

r = the correlation between the two half tests (even numbered and odd numbered items.

N = Total number of scores.

$\sum X$ = Sum of scores from the first half test (even numbered items).

$\sum \tilde{Y}$ = Sum of scores from the second half test (odd numbered items).

$\sum X^2$ = Sum of the squared scores from the first half test.

$\sum \tilde{Y}^2$ = Sum of the squared scores from the second half test.

$\sum X\tilde{Y}$ = Sum of the product of the scores from the first and the second half tests.

The Spearman - Brown prophecy formula was used to adjust the correlation coefficient (r) obtained, to reflect the correlation coefficient of the full-length test (Mozube, 1987; Gall and Borg, 1996; Gay, 1987).

$R = \frac{2r}{1+r}$ Where: R= Estimated reliability of the full-length test.
r = the actual correlation between the two half-length tests.

The standard error of measurement (SEM) was determined using the formula given below.

$$SEM = SD\sqrt{1-r}$$

Where SEM = Standard error of measurement.
SD = the standard deviation of the test scores.
r = the reliability coefficient.

Secondly, the alternative form reliability was determined, whereby, scores from the developed test and those from TIPS were correlated using the Pearson product-moment coefficient as shown above.

3.7.6 READABILITY OF THE INSTRUMENT

The Flesch reading ease formulae was used to determine the readability of the test instrument. The computation of the readability level was based on 15 items sampled randomly from the items in the developed instrument. Words and sentences associated with graphs, charts, and tables were excluded from the texts that were used in the computation of this index. To determine the readability level, the following steps were carried out;

1. The average sentence length (ASL) was determined (ie. Number of words per sentence).

2. The average number of syllables per word (ASW) was determined.

3. The readability score of the instrument was estimated by substituting ASL and ASW, in the following Flesch reading ease formula:

   Readability score = 206.835 – (1.015 x ASL) – (84.6 x ASW)

4. Interpretation of the Flesch reading ease scores

The following scale was used to estimate the level of reading difficulty of the developed test, using the score obtained from the Flesch reading ease formula.

Readability score

100	Very easy
90	Easy
80	Fairly easy
70	**Plain English**
60	Fairly difficult
50	↓
40	Difficult
30	↓
20	Very difficult
10	↓
0	↓

The higher the readability score, the easier the text is to understand, and vice versa. The recommended range of scores for a test instrument is 60 to 70, which is the plain English level (Klare, 1976). The results from the reading ease scale showed that the developed test instrument had a fairly easy readability.

3.7.6.1 READING GRADE LEVEL OF THE DEVELOPED INSTRUMENT

The reading grade level is the value that is determined to estimate the grade level for which a given text is suitable (Klare, 1976). For example, a score of 10, means that a tenth grader (in the European context) would understand the text easily (Klare, 1976). In this study, the Flesh-Kincaid formula (Klare, 1976) was used to make an approximation of the appropriate reading grade (school age) level of the developed instrument. The Flesch-Kincaid formula is shown below.

Grade level score = (0.39*ASL) + (11.8*ASW) – 15.9

Where;
ASL = Average sentence length
ASW = Average number of syllables per word

3.7.7 COMPARISON OF THE PERFORMANCE OF LEARNERS FROM DIFFERENT GROUPS.

The means of the different samples were compared, and the significance of any differences observed between the groups, such as between; urban and rural schools, white and black learners, and girls and boys, were determined using the t-test, as indicated below. The significance of any difference observed between the learners' performance on the developed test and TIPS was also determined using the t-test for paired samples. The comparison of the performance of the learners from the different school types, and different grades involved three variables (formerly model C and DET schools, and private schools; and grades 9, 10 and 11). The simple ANOVA was therefore used to determine the significance of the differences observed among the means of these variables.

The formulae used for these computations are shown below**.**

The t-test for independent samples (Gay, 1987)

For the null hypothesis Ho; $\mu 1 = \mu_2$

and the alternative hypothesis H_a; $\mu_1 \neq \mu_2$.

F_{cv} at $\alpha = 0.05$

$$T = \frac{X_1 - X_2}{\sqrt{\left(\frac{SS_1 + SS_2}{n_1 + n_2 - 2}\right)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Where ; X_1 = mean of sample 1 X_2 = mean of sample 2

n_1 = number of learners in sample 1.

n_2 = number of learners in sample 2.

SS_1 = sum of squares for sample 1.

SS_2 = sum of squares for sample 2.

One-way analysis of variance (ANOVA) (Gay, 1987)

For the null hypothesis H_o; $\mu_1 = \mu_2 = \mu_3$

and the alternative hypothesis H_a; $\mu_1 \neq \mu_2$, for some $_{i,k}$.; F_{cv} at $\alpha = 0.05$

TABLE 3.10 ONE WAY ANALYSIS OF VARIANCE (ANOVA)

Source of Variation	Sum of squares-SS	Degree of freedom-df	Mean Square-MS	F- ratio	(F_{cv})
Between	$\sum n_K(X_k - X)^2$	**K -1**	**SSB/K - 1**	**MS_B/MS_W**	
Within	$\sum\sum(Xi_k - X_k)^2$	**N - K**	**SS_W/N - K**		
Total	$\sum\sum(Xi_k - X)^2$	**N -1**			

Where; X = Grand mean X_k = Sample mean

X_{ik} = The i^{th} score in the k^{th} group K = Number of groups

N = Total sample size SS_B = Between sum of squares

SS_W = Within sum of squares MS_B = Between mean square

MS_W = Within mean square

3.8 ETHICAL ISSUES

The participants were duly informed of the objectives of the study before the test was administered to them. All the procedures that involved the participants were explained to them, and they were informed of their right to decline from participating in the study, if they so wished. The participants were given number codes, to ensure that they remain anonymous to external populations. The test scripts were handled by the researcher and her assistants only. The scripts were stored in a safe place, after marking, and they will be destroyed three years after the study. The performance of each school on the test is highly confidential. Participating schools were promised access to their results on request. The study report will be submitted to the supervisor of the study, the Limpopo Department of Education, and possibly be presented at a Southern African Association for Research in Mathematics, Science, and Technology Education (SAARMSTE) conference, or other similar conferences. The researcher also intends to publish the results of the study.

CHAPTER 4
RESULTS AND DISCUSSION

This chapter analyses and discusses the results of the study. The statistical procedures outlined in section 3.7 were used for data analysis. The results are presented in the following order: the item response pattern, discrimination index, index of difficulty, reliability, readability level of the instrument, and the comparison of the performance of different groups of learners on the developed test.

4.1 ITEM RESPONSE PATTERN

Scores from all the 769 learners involved in the study were used to determine the item response pattern. The learners were divided into performance categories (ie. high, medium and low scorers), as described in section 3.7.2. The maximum score obtained in the main study was 100%, while the minimum score was 7%. The item response pattern was organized according to the performance categories, the different grade levels, and the integrated science process skills measured, as explained in the following texts.

4.1.1 ITEM RESPONSE PATTERN ACCORDING TO PERFORMANCE CATEGORIES.

The item response pattern for all the learners who participated in the study was determined according to their performance categories (high, medium, and low scorers). The percentages of learners who selected each option in the different performance categories are shown in table 4.1 below. Detailed information on the item response pattern according to performance categories is given on Appendices III and IV.

As evident from table 4.1, each distracter was selected by a sufficient number (more than 2% of the total number of subjects) of learners from all the three performance categories. The distracters may therefore be considered to be plausible.

TABLE 4.1 PERCENTAGE OF LEARNERS WHO SELECTED EACH OPTION IN EACH PERFORMANCE CATEGORY.

Qn #	Option A			Option B			Option C			Option D			Others		
	H	M	L	H	M	L	H	M	L	H	M	L	H	M	L
1	3.8	10	13	9.6	31	34	4.8	13	24	**75**	**42**	**38**	0.5	1.1	0.5
2	5.8	22	26	**81**	**63**	**40**	3.4	6.8	19	9.6	8.5	15	0	0.3	0
3	10	7.1	25	12	19	23	10	16	18	**75**	**54**	**31**	0	1.4	1
4	4.3	16	29	**78**	**51**	**38**	9.1	18	25	6.3	12	11	1	0.8	1.9
5	**63**	**27**	**18**	33	59	66	4.3	6.5	9.1	3.8	4	6.3	0	0.8	1.4
6	1.9	5.9	15	12	25	29	**60**	**32**	**33**	21	33	30	1	0.8	1
7	16	14	19	**50**	**29**	**18**	23	25	30	10	27	37	0.5	2	1
8	22	27	19	8.2	11	22	3.8	7.1	15	**65**	**55**	**41**	0.5	0.6	1.9
9	3.4	3.1	14	**73**	**61**	**42**	15	30	34	5.8	4	14	0	0.6	0.5
10	**72**	**40**	**21**	3.8	8.8	16	19	38	52	4.3	8.8	15	1	1.1	1.9
11	6.3	20	23	**52**	**20**	**15**	10	24	27	32	35	35	0	0.8	1
12	**66**	**41**	**23**	2.9	14	19	1.4	15	26	27	31	32	0	0.3	1
13	3.8	7.1	25	2.9	15	24	**77**	**50**	**45**	11	24	18	0	0.3	0
14	31	36	35	**47**	**29**	**19**	13	20	26	5.3	16	22	0	0.8	0
15	20	28	34	15	16	29	17	20	23	**47**	**34**	**13**	1	0.3	2.4
16	5.3	14	29	21	22	31	25	40	39	**45**	**19**	**13**	0	0.3	1
17	7.2	18	27	**58**	**31**	**20**	11	17	24	24	31	33	0.5	0.6	1
18	6.3	24	34	9.6	22	25	**70**	**41**	**39**	4.3	7.1	19	0.5	0.8	1.4
19	15	24	27	9.6	22	22	3.4	16	33	**67**	**38**	**20**	0	0.6	1.9
20	6.3	19	18	3.8	13	18	**76**	**32**	**34**	4.8	35	39	1	0.3	1.4
21	22	30	32	3.8	22	24	**55**	**26**	**25**	13	22	23	0	1.4	1
22	**59**	**31**	**19**	20	40	29	7.7	9.9	13	11	25	30	0	0.8	1
23	23	24	29	21	29	39	9.6	20	19	**46**	**25**	**14**	1	1.1	1
24	**61**	**50**	**34**	20	25	23	9.1	13	25	10	8.2	24	0.5	1.7	0.5
25	18	17	25	13	22	24	**61**	**41**	**41**	7.2	14	18	1	0.8	1.4
26	**85**	**52**	**26**	9.6	21	31	4.8	17	27	1.9	6.2	9.6	1.4	2.5	5.3
27	**44**	**18**	**15**	18	28	37	20	21	29	17	27	25	1	1.1	2.4
28	12	23	28	8.7	16	26	13	15	23	**66**	**41**	**23**	1	2.8	1.9
29	31	34	43	5.3	13	17	**48**	**29**	**33**	7.2	17	21	1.9	2	1
30	4.3	14	23	**68**	**35**	**17**	21	31	42	5.3	14	23	1.9	2.3	1.9
31	13	25	24	3.8	18	19	**63**	**26**	**35**	14	26	32	1	0.6	1.4

KEY:

Qn # = item number; Bold (red) = correct option.

H = Percentage of high scorers who selected the option.

M = Percentage of medium scorers who selected the option.

L = Percentage of low scorers who selected the option.

Others = Percentage of learners who omitted the item or selected more than one option.

Table 4.1 also shows that, for almost all the items, a higher percentage of the high scorers selected the correct option, followed by that of the medium scorers, while a lower percentage of low scorers selected the correct option. For instance, from table 4.1, item number 2, 81% of the high scorers selected the correct option, 63% of the medium scorers selected the correct option, and 40% of the low scorers selected the correct option Conversely, the distracters attracted more of the low scorers and fewer high scorers. For example, from table 4.1, item number 1, option C (a distracter), 24% of the low scorers selected it, 13% of the medium scorers selected it, and only 4.8% of the high scorers selected it. Refer to Appendices III and IV for more details on the item response pattern. These results suggest that the developed test was able to discriminate between those who are likely to be more competent in science process skills (high scorers) and those who are likely to be less competent in the skills (low scorers).

4.1.2 ITEM RESPONSE PATTERN ACCORDING TO GRADE LEVELS.

The item response pattern was also arranged according to the grade levels of the learners that participated in the study (Table 4.2 and Appendices VI).

TABLE 4.2 PERCENTAGE OF LEARNERS SELECTING THE CORRECT OPTION FOR EACH ITEM, ACCORDING TO GRADE LEVELS AND PERFORMANCE CATEGORIES

Scoring group	Item Number	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	**Ave %**
	Grade																																
High	9	80	83	72	80	61	51	44	54	75	62	51	54	70	45	30	42	49	56	54	56	55	42	31	59	54	80	38	52	28	54	51	**55**
scorers	10	72	84	74	78	62	59	49	70	68	62	52	71	83	49	45	43	55	75	77	94	45	81	43	48	57	86	49	68	55	65	46	**63**
(in %)	11	74	76	78	75	66	71	59	72	75	93	53	75	79	46	68	49	69	79	72	79	65	54	63	75	74	88	46	78	60	85	87	**70**
Medium	9	52	58	34	49	38	30	29	49	56	31	16	35	39	25	23	16	28	34	34	21	32	11	18	51	41	48	25	32	25	30	14	**33**
scorers	10	48	66	59	57	32	32	32	54	60	33	22	39	59	26	26	26	36	33	44	32	21	38	24	44	41	47	17	38	24	31	21	**37**
(in %)	11	26	65	68	47	20	34	25	61	68	54	20	45	52	36	54	16	28	55	36	43	25	18	35	64	40	63	20	54	39	45	45	**42**
Low	9	45	44	27	41	10	18	23	38	34	13	11	24	20	13	7	11	21	15	14	13	17	11	3	28	32	25	14	11	13	7	15	**20**
scorers	10	35	45	30	38	28	25	14	39	43	17	14	26	30	19	14	12	19	12	19	17	14	32	14	28	25	30	19	33	22	16	22	**24**
(in %)	11	35	31	37	37	16	31	18	47	49	32	19	19	32	25	19	15	19	21	26	21	19	13	25	46	21	22	13	25	18	29	26	**26**

Ave % = Average % per grade

The results from this analysis indicated that, in all grades, the correct options attracted more of the high scorers than the others (Table 4.2). For instance, for item number 8, the percentages of high, medium and low scorers who selected the correct option in grade 9 were 54%, 49%/ and 38%, while in grade 10 were 70%, 54, and 39, and in grade 11 were 72%, 61%, and 47% respectively (Table 4.2). This trend can also be seen from the average percentages, as shown on table 4.2.

The average percentages further show that, more of the grade 11 learners selected the correct options, followed by the grade 10 learners and then the grade 9 learners, in all the performance categories (Table 4.2). For example, in the high scoring category, 70% of grade 11 learners selected the correct options, 63% of grade 10 learners selected the correct option, and lastly 55% of grade 9 learners selected the correct options. These results suggest that learners in lower grades found the test to be more difficult than learners in higher grades. This implies that the developed instrument can discriminate well between learners who have more experience in activities involving science process skills (higher grade levels) and those who do not have (lower grade levels).

4.1.3 ITEM RESPONSE PATTERN ACCORDING TO THE PROCESS SKILLS MEASURED.

The item response pattern was further arranged according to the science process skills measured (Table 4.3). The table shows how the learners from the different performance categories performed in each science process skill considered.

TABLE 4.3 PERCENTAGE OF LEARNERS WHO SELECTED THE CORRECT OPTION FOR ITEMS RELATED TO EACH SCIENCE PROCESS SKILL TESTED FOR.

Science Process skill	Item Numbers	High scorers (%)	Medium scorers (%)	Low scorers (%)
Identifying and controlling variables	2,6,20,26,29,30,31	69	38	31
Stating hypotheses	9,13,17,21,24,27	61	39	30
Operational definitions	1,7,11,19,22,23	58	31	21
Graphing and interpreting data	4,5,8,10,12,15,18,25,28	65	41	29
Experimental design	3,14,16	56	34	21

The results show that more of the high scorers (69%) selected the correct options on items related to the science process skill of identifying and controlling variables than other skills. This skill is followed by the skill of graphing and interpretation of data, where 65% of the high scorers selected the correct options on items related to it. Items related to the skill of operational definitions had a smaller percentage of high scorers who selected the correct option (58%). Items related to the skill of designing experiments attracted the least percentage of high scorers, whereby only (56%) selected the correct options. This trend was more or less the same for all the performance categories. See Appendix V, for detailed information on this pattern. This result suggests that the learners involved in the study were less competent in the skill of designing investigations.

The item response pattern of the different process skills was further arranged according to grade levels and performance categories, to show how learners from the different performance categories in each grade responded (Table 4.4).

TABLE 4.4 PERCENTAGE OF LEARNERS SELECTING THE CORRECT OPTION FOR EACH PROCESS SKILL , ACCORDING TO GRADE LEVELS AND PERFORMANCE CATEGORIES

		High scorers (%)			Medium scorers (%)			Low scorers (%)		
SCIENCE PROCESS SKILL	**Item numbers**	**9**	**10**	**11**	**9**	**10**	**11**	**9**	**10**	**11**
Identifying and controlling variables	2,6,20,26,29,30,31	58	70	78	32	36	48	19	25	27
Stating hypotheses	9,13,17,21,24,27	58	58	71	39	40	43	22	26	30
Operational definitions	1,7,11,19,22,23	50	62	63	27	35	43	18	21	22
Graphing and interpreting data	4,5,8,10,12,15,18,25,28	51	65	76	37	39	48	21	26	28
Experimental design	3,14,16	53	55	58	25	37	49	17	20	26

The results from the above table show that first, in each performance category, more grade 11 learners selected the correct options, followed by grade 10 learners, and fewer grade 9 learners (Table 4.4), further highlighting the discriminatory power of the developed test.

Second, more learners from the different performance groups in each grade selected the correct options for items related to the skill of identifying and controlling variables. In other words, learners from the different grade levels found the skill of identifying and controlling variables easier than other skills (Table 4.4). While fewer learners from the different performance categories in each grade selected the correct option for items related to the skill of designing experiments (Table 4.4). Suggesting the possibility of learners having less experience in designing experiments, and the likelihood of the use of prescribed experimental designs, in science classes.

Thirdly, at each grade level, more learners from the high scoring group selected the correct options, for items related to each process skill, than those from the medium and low scoring groups (Table 4.4). Few learners from the low scoring group selected the correct options on items related to each processes skill (Table 4.4).

This result also shows that the test instrument is able to discriminate between learners who are competent in science process skills (high scorers) and those who are not (low scorers).

4.2 DISCRIMINATION INDICES

The discrimination indices of the items were organized according to the different grade levels and the integrated science process skills measured.

4.2.1 DISCRIMINATION INDICES ACCORDING TO GRADE LEVELS

The discrimination indices of the test items were determined according to grade levels, in order to find out the discrimination power of the developed test instrument in the different grade levels.

The discrimination index for each item was determined using the scores of the high scorers and low scorers as discussed in section 3.7.3. The results are presented on table 4.5. The table shows that, the values of the discrimination indices increase as the grade levels increase, [ie. grade 9 = 0.36, grade 10 = 0.40, grade 11 = 0.45] (Table 4.5). This suggests that the instrument discriminated better among learners in the higher grade levels than those in the lower levels. The overall discrimination index of the instrument was 0.40 (Table 4.5). This value is well within the recommended range of values for this test characteristic (ie $\geq$ 0.3).

Further analysis of table 4.5 shows that, 13% of the items had discrimination indices of less than 0.3. However, 3 of the 4 items in this category had discrimination indices which were very close to 0.3. These items were therefore retained in the test. Forty two percent of the items had discrimination indices that fell between 0.3 and 0.4, 26% had discrimination indices that fell between 0.4 and 0.5, while 19% of the items had discrimination indices of more than 0.5 (See Appendix VII for detailed information). Of the 31 items analyzed, only item 8 had a very low discrimination index (0.24). It was therefore necessary to discard this item.

TABLE 4.5 DISCRIMINATION INDICES FOR EACH ITEM ACCORDING TO GRADES.

Item		DISCRIMINATION INDEX			
No.	PROCESS SKILL MEASURED	Grade 9	Grade 10	GRD 11	OVER-ALL
1	Operational definitions	0.38	0.38	0.38	**0.38**
2	Identifying and controlling variables	0.39	0.39	0.46	**0.41**
3	Experimental design	0.45	0.43	0.41	**0.43**
4	Graphing and interpreting data	0.39	0.41	0.38	**0.39**
5	Graphing and interpreting data	0.50	0.35	0.5	**0.45**
6	Identifying and controlling variables	0.32	0.35	0.34	**0.36**
7	Operational definitions	0.21	0.35	0.41	**0.32**
***8**	Graphing and interpreting data	0.15	0.30	0.25	**0.24**
9	Stating hypotheses	0.41	0.25	0.26	**0.31**
10	Graphing and interpreting data	0.49	0.45	0.60	**0.52**
11	Operational definitions	0.39	0.38	0.34	**0.37**
12	Graphing and interpreting data	0.30	0.45	0.56	**0.43**
13	Stating hypotheses	0.51	0.52	0.47	**0.5**
14	Experimental design	0.32	0.30	0.21	**0.28**
15	Graphing and interpreting data	0.22	0.30	0.49	**0.34**
16	Experimental design	0.31	0.32	0.34	**0.32**
17	Stating hypotheses	0.28	0.36	0.5	**0.38**
18	Graphing and interpreting data	0.41	0.64	0.59	**0.54**
19	Operational definitions	0.39	0.58	0.46	**0.48**
20	Identifying and controlling variables	0.44	0.77	0.59	**0.6**
21	Stating hypotheses	0.38	0.30	0.46	**0.38**
22	Operational definitions	0.31	0.49	0.41	**0.40**
23	Operational definitions	0.28	0.29	0.38	**0.32**
24	Stating hypotheses	0.31	0.20	0.29	**0.27**
25	Graphing and interpreting data	0.21	0.32	0.53	**0.35**
26	Identifying and controlling variables	0.55	0.55	0.66	**0.59**
27	Stating hypotheses	0.24	0.30	0.32	**0.29**
28	Graphing and interpreting data	0.41	0.35	0.53	**0.43**
29	Identifying and controlling variables	0.15	0.33	0.43	**0.30**
30	Identifying and controlling variables	0.46	0.49	0.55	**0.50**
31	Identifying and controlling variables	0.35	0.39	0.57	**0.44**
X		**0.35**	**0.39**	**0.44**	**0.40**
X*	Averages after eliminating item 8	**0.36**	**0.4**	**0.45**	**0.40**

4.2.2 DISCRIMINATION INDICES ACCORDING TO THE PROCESS SKILLS MEASURED.

The discrimination indices of the items were further grouped according to the science process skills measured in the study. This was necessary to determine the science process skills which discriminated better than others. The results of this analysis are shown on table 4.6 below.

Analysis of the results show that the items related to the skill of identifying and controlling variables had the highest discrimination power, with an average discrimination index (D) of 0.46, followed by that of the items related to the skill of graphing and interpreting data (D = 0.43). The items related to the skill of stating hypotheses had a low discriminating power (D = 0.35), and those related to the skill of designing experiments had the lowest discrimination power (D = 0.34). However, all these indices fall within the acceptable range of values for this test characteristic (0.3 – 0.1). See table 4.6 for the cited discrimination indices.

TABLE 4.6. DISCRIMINATION INDICES ACCORDING TO THE SCIENCE PROCESS SKILLS MEASURED.

Key:

Obje # = The number of the object to which the item is referenced.

Item # = The number of the item in the test instrument.

Discrimina = Discrimination index.

			Item #		Obje. #	Discrimination
A	Identifying and controlling variables				**1 and 3**	
			2		1	**0.41**
			6		**3**	**0.36**
			19		**3**	**0.60**
			25		**3**	**0.59**
			28		**1**	**0.30**
			29		**1**	**0.50**
			30		**1**	**0.44**
			Average			**0.46**
B	Stating hypotheses				**4 and 6**	
			8		**6**	**0.31**
			12		**6**	**0.50**
			16		**6**	**0.38**
			20		**4**	**0.38**
			23		**4**	**0.27**
			26		**4**	**0.29**
			Average			**0.35**
C	Operational definitions				**2 and 5**	
			1		**5**	**0.38**
			7		**2**	**0.32**
			10		**5**	**0.37**
			18		**2**	**0.48**
			21		**2**	**0.40**
			22		**5**	**0.32**
			Average			**0.38**
D	Graphing and interpreting data				**8 and 9**	
			4		**9**	**0.39**
			5		**9**	**0.45**
			9		**8**	**0.52**
			11		**9**	**0.43**
			14		**8**	**0.34**
			17		**9**	**0.54**
			24		**8**	**0.35**
			27		**9**	**0.43**
			Average			**0.43**
E	Experimental design				**7**	
			3		**7**	**0.43**
			13		**7**	**0.28**
			15		**7**	**0.32**
			Average			**0.34**

4.3 INDICES OF DIFFICULTY

The indices of difficulty of the items were organized according to the different grade levels and the integrated science process skills measured.

4.3.1 INDICES OF DIFFICULTY ACCORDING TO GRADE LEVELS

The values of the indices of difficulty for the different grade levels also increase as the grades increase [grade 9 = 0.35, grade 10 = 0.40, grade 11 = 0.45] (Table 4.7). In this case, the increase in the indices of difficulty suggests that the learners from higher grades found the test to be easier than those in the lower grades. This result is expected, since learners in higher grades are expected to be more experienced with activities involving science process skills than those in lower grades. The above indices all fall within the acceptable range of values for indices of difficulty [0.4 - 0.6], (Nitko, 1996).

Table 4.7 shows that thirteen percent of the items had indices of difficulty of less than 0.3, and these, according to literature are considered to be difficult (Nitko, 1996). Thirty five percent of the items had indices of difficulty that fell between 0.3 and 0.4, which are also considered to be difficult. Twenty six percent of the items had indices of difficulty that fell between 0.4 and 0.5. Twenty three percent of them had indices of difficult that fell between 0.5 and 0.6. Items that fell within the latter two ranges are considered to be fair. Hence 49% of the items are fair. Three percent of the items had indices of difficulty of more than 0.6. These items are considered easy. Specifically, items 5, 6, 7, 11, 14, 15, 16, 17, 21, 22, 23, 27 and 31 had low indices of difficulty, of less than 0.4 (Table 4.7). These items are therefore considered to be difficult. As a result, the overall index of difficulty was quite low (0.40), indicating that the learners may have found the test to be generally difficult. However, these items were retained in the instrument despite the low indices of difficulty, because they had good discrimination indices. In other words, they were able to discriminate between learners who are competent in integrated science process skills and those who are not.

TABLE 4.7 INDICES OF DIFFICULTY FOR EACH ITEM ACCORDING TO GRADES.

Item		**INDICES OF DIFFICULTY**			
NO.	PROCESS SKILL MEASURED	Grade 9	Grade 10	Grade 11	OVERALL
1	Operational definitions	0.58	0.51	0.42	**0.50**
2	Identifying and controlling variables	0.61	0.65	0.59	**0.62**
3	Experimental design	0.42	0.55	0.62	**0.53**
4	Graphing and interpreting data	0.55	0.58	0.52	**0.55**
5	Graphing and interpreting data	0.36	0.39	0.32	**0.36**
6	Identifying and controlling variables	0.33	0.37	0.43	**0.38**
7	Operational definitions	0.31	0.32	0.32	**0.32**
***8**	Graphing and interpreting data	0.47	0.54	0.60	**0.54**
9	Stating hypotheses	0.55	0.58	0.64	**0.59**
10	Graphing and interpreting data	0.34	0.37	0.59	**0.43**
11	Operational definitions	0.24	0.28	0.29	**0.27**
12	Graphing and interpreting data	0.37	0.44	0.46	**0.42**
13	Stating hypotheses	0.42	0.58	0.54	**0.51**
14	Experimental design	0.27	0.30	0.36	**0.31**
15	Graphing and interpreting data	0.20	0.28	0.48	**0.32**
16	Experimental design	0.22	0.27	0.24	**0.24**
17	Stating hypotheses	0.32	0.36	0.37	**0.35**
18	Graphing and interpreting data	0.35	0.39	0.52	**0.42**
19	Operational definitions	0.34	0.46	0.43	**0.41**
20	Identifying and controlling variables	0.28	0.45	0.47	**0.40**
21	Stating hypotheses	0.34	0.26	0.34	**0.31**
22	Operational definitions	0.20	0.48	0.26	**0.31**
23	Operational definitions	0.17	0.27	0.4	**0.28**
24	Stating hypotheses	0.47	0.40	0.6	**0.50**
25	Graphing and interpreting data	0.42	0.41	0.44	**0.42**
26	Identifying and controlling variables	0.51	0.53	0.59	**0.54**
27	Stating hypotheses	0.25	0.26	0.25	**0.26**
28	Graphing and interpreting data	0.32	0.45	0.53	**0.43**
29	Identifying and controlling variables	0.23	0.32	0.39	**0.31**
30	Identifying and controlling variables	0.30	0.36	0.52	**0.39**
31	Identifying and controlling variables	0.24	0.28	0.52	**0.35**
X		**0.36**	**0.41**	**0.45**	**0.41**
X*	Averages after eliminating item 8	**0.35**	**0.40**	**0.45**	**0.40**

4.3.2 INDICES OF DIFFICULTY ACCORDING TO THE SCIENCE PROCESS SKILLS MEASURED.

The indices of difficulty of the items were further grouped according to the science process skills measured. This was necessary to identify the process skills which the learners found to be more difficult than others. The results of this analysis are shown in table 4.8 below.

Data from table 4.8 suggests that, learners found the items related to the skill of making operational definitions and those related to the skill of designing experiments (average difficulty indices of 0.35 and 0.36 respectively) to be more difficult than those related to the other skills considered in this study, which had average indices of difficulty of about 0.42 (Table 4.8).

The low value of the indices of difficulty for the skills of designing experiments, and making operational definitions further shows that, the learners involved in this study found the items related to these two skills difficult.

TABLE 4.8. INDICES OF DIFFICULTY ACCORDING TO THE SCIENCE PROCESS SKILLS MEASURED.

Key:

Obje # = The number of the object to which the item is referenced.

Item # = The number of the item in the test instrument.

Difficulty = Index of difficulty.

			Item #		Obje. #	Difficulty
A	Identifying and controlling variables				**1 and 3**	
			2		1	**0.61**
			6		**3**	**0.38**
			19		**3**	**0.40**
			25		**3**	**0.54**
			28		**1**	**0.31**
			29		**1**	**0.39**
			30		**1**	**0.35**
			Average			**0.43**
B	Stating hypotheses				**4 and 6**	
			8		**6**	**0.59**
			12		**6**	**0.51**
			16		**6**	**0.35**
			20		**4**	**0.31**
			23		**4**	**0.50**
			26		**4**	**0.26**
			Average			**0.42**
C	Operational definitions				**2 and 5**	
			1		**5**	**0.50**
			7		**2**	**0.32**
			10		**5**	**0.27**
			18		**2**	**0.41**
			21		**2**	**0.31**
			22		**5**	**0.28**
			Average			**0.35**
D	Graphing and interpreting data				**8 and 9**	
			4		**9**	**0.55**
			5		**9**	**0.36**
			9		**8**	**0.43**
			11		**9**	**0.42**
			14		**8**	**0.32**
			17		**9**	**0.42**
			24		**8**	**0.42**
			27		**9**	**0.43**
			Average			**0.42**
E	Experimental design				**7**	
			3		**7**	**0.53**
			13		**7**	**0.31**
			15		**7**	**0.24**
			Average			**0.36**

4.4 RELIABILITY OF THE TEST INSTRUMENT

The reliability of the developed instrument was estimated using the split half method of determining the internal consistency reliability, the standard error of measurement, and the alternative form reliability. These coefficients were determined as explained in section 3.7.5.

4.4.1 INTERNAL CONSISTENCY RELIABILITY

The correlation coefficients (r) for the internal consistency reliability on the half tests (odd and even numbered items), in the different grade levels were: 0.683 for grade 9; 0.67 for grade 10; and 0.681 for grade 11 (See appendix VIII and XII). The Spearman - Brown prophecy formula was used to adjust these correlation coefficients (r) for the half tests, to reflect the correlation coefficient of the full-length test, as follows:

$$R = \frac{2r}{1+r}$$

$$R = \frac{2 * 0.683}{1 + 0.683} = \mathbf{0.811} \text{ for grade 9}$$

$$R = \frac{2 * 0.67}{1 + 0.67} = \mathbf{0.802} \text{ for grade 10}$$

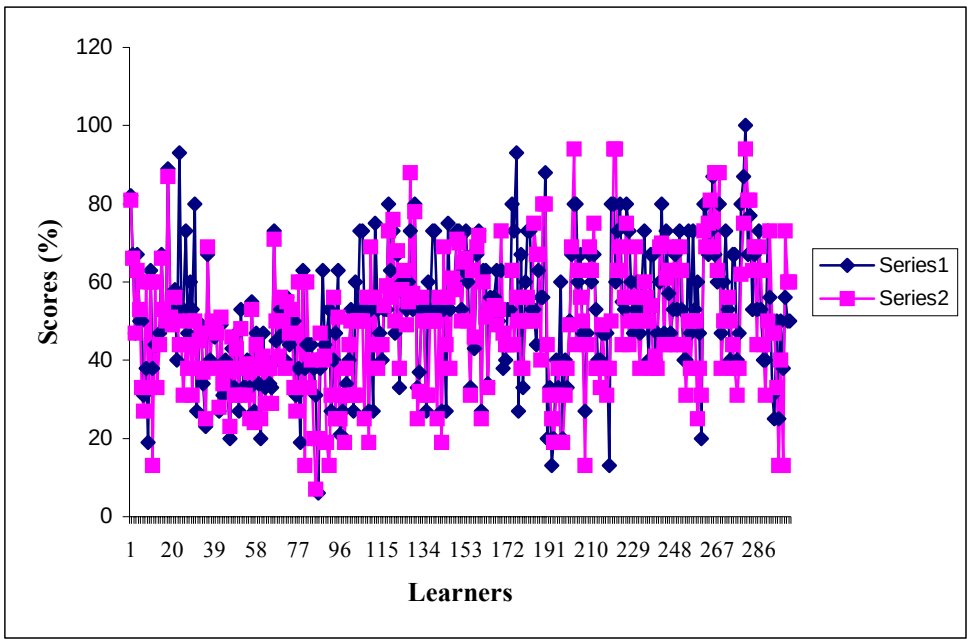
$$R = \frac{2 * 0.681}{1 + 0.681} = \mathbf{0.810} \text{ for grade 11}$$

Overall reliability $R = 0.811 + 0.802 + 0.810 = \mathbf{0.808} = \underline{\mathbf{0.81}}$

This reliability coefficient is well above the lower limit of the acceptable range of values for reliability [0.70 – 1.0] (Adkins, 1974; Hinkle, 1998), and it is within the range of reliability coefficients obtained from similar studies, such as; Dillashaw and Okey (1980) who obtained a reliability of 0.89, Onwu and Mozube (1992) who obtained a reliability of 0.84, and Molitor and George (1976) who obtained reliabilities of 0.77 and 0.66 for skills of inference and verification respectively. The developed test may therefore be considered reliable. The final reliability of the test instrument (0.81), is an improvement from the reliability obtained from the pilot study (0.73).

Figure 4.1 shows a fair distribution of the scores from the even and odd numbered items of the instrument. That is, the learners' performance on both half tests is equally distributed, affirming the fact that the two half tests had the same effect on the learners.

FIG. 4.1. GRAPH COMPARING SCORES FROM EVEN AND ODD NUMBERED ITEMS OF THE INSTRUMENT



Series 1. Scores from even-numbered items ⟶ Mean score = 51.99

Series 2. Scores from odd-numbered items ⟶ Mean score = 50.57

The graph on Appendix XII shows a positive correlation between the scores obtained from even and odd numbered items of the test instrument. This further shows that the performance of the learners on the even and odd numbered items of the test instrument was similar. (See appendix VIII for scores obtained by subjects on the even and odd-numbered items).

4.4.2. STANDARD ERROR OF MEASUREMENT

The formula discussed in section 3.7.5 was used to determine the Standard Error of Measurements (SEM), to further estimate the reliability of the instrument. The standard error of measurement was determined as follows:

SEM $= SD\sqrt{1 - r}$ Where: SD = standard deviation

SEM $= 16.12\sqrt{1 - 0.8078}$ r = reliability coefficient

$= 16.12 * 0.438406$

$= 7.0671$

This value (7.07) is relatively small, which means that the learners' obtained scores did not deviate much from their true scores. The smaller the standard error of measurement, the more reliable the results will be (Nitko, 1996).

4.4.3 ALTERNATIVE FORM RELIABILTY

The alternative form reliability was obtained by correlating the learners' scores obtained from the developed test, and from the TIPS (Dillashaw and Okey, 1980). The data used in this computation were from the school where the two tests were administered concurrently. The correlation coefficient obtained was 0.56. This value is below the acceptable range of value for reliability (≤ 0.7). The determination of this coefficient involved the use of the TIPS, which as explained in sections 1.1, 2.5.4.1, 2.5.42, and 4.5.5, was not suitable for use in this specific case. This correlation was nevertheless necessary to show that local learners performed differently on the two tests. The alternative form reliability was therefore was not considered in the determination of the reliability of the developed test.

4.5 READABILITY LEVEL OF THE DEVELOPED INSTRUMENT

The readability of the final test instrument was obtained using the Flesch reading ease formula as outlined in section 3.7.6. The Flesch reading ease scale is rated from 0 to 100. A high readability value implies an easy to read text. The suggested range for a fairly easy readability level is 60 to 70 (Klare, 1976).

The readability level of the developed instrument was found to be 70.29 (see below). This readability level is on the higher end of the 'fairly easy readability range,' on Flesch's reading ease scale. Therefore, the readability level of the developed instrument may be considered fairly easy. The calculation of the readability level was done as shown below. The data used to calculate the readability level of the developed test instrument is shown in Appendix IX.

The average sentence length (ASL) = **15.95** and
The average number of syllables per word (ASW) = **1.42**

Readability score **= (206.835 - (1.015*ASL) - (84.6*ASW)**
= (206.835 - (1.015*15.95) - (84.6*1.42)
= 70.29

4.5.1 READING GRADE LEVEL OF THE DEVELOPED INSTRUMENT

The results obtained from the calculation of the reading grade level for the developed instrument showed that, the suitable reading level of the developed instrument is grade 8. This value was determined manually, as shown below, as well as by using a computer program (Microsoft word 2000).

It is pertinent to point out that, the determination of the Flesh-Kincaid formula was based on grade levels from schools in European countries, where English is used as a first language. For most South African learners, English is used as a second or third language. The actual grade levels for South African users of the test instrument is therefore likely to be higher than that suggested by the formula. This argument is supported by Stephens (2000) who states that at higher-grade levels, grade level scores are not reliable, because, background and content knowledge become more significant than style variables. They (grade levels) are therefore likely to under-estimate or over-estimate the suitability of the material.

Furthermore, one of the arguments raised in this study is that, the language used in the tests of science process skills developed outside South Africa, put South African users of such tests at a disadvantage. A test developed for South African learners should therefore have a fairly easier readability level, than one developed and validated for first language English users.

Calculation of the grade level score of the test instrument, using the Flesch-Kincaid formula

The average sentence length **(ASL) = 15.95349 and**

The average number of syllables per word **(ASW) = 1.422565**

Grade level score =(0.39*ASL) + (11.8*ASW) – 15.9
=(0.39*15.95349) + (11.8*1.422565) – 15.9
=7.418 approximated to 8

Key: ASL = Average sentence length.
ASW = Average number of syllables per word.

Computer result: Grade level score = 8

Given the above arguments, a reading grade level of 8, suggests that the test text is likely to be easy to read by the target population (further education and training learners).

4.6 COMPARISON OF THE PERFORMANCE OF DIFFERENT GROUPS OF LEARNERS

Steps were taken during the development and validation of the test instrument, to assure that the test instrument was not biased against some groups of learners (section 3.3.1). The performances of the different groups of learners (gender, location, school type, and grade level) who participated in the study were compared, to get an indication of whether the test was biased against some groups of learners or not. The following passages describe the results of these comparisons.

One of the assumptions made about the data collected in this study is that it is a normal distribution. In consequence, parametric tests (t-test and ANOVA) were used to determine whether any mean differences observed among any set of data were significantly different or not, as shown on tables 4.9 to 4.14.

The number of subjects (N) used in each category was determined by identifying the group with the smallest number of subjects among the groups involved in the category. This (smallest) number of subjects was randomly selected from the groups with larger number of subjects, to obtain the same number of subjects per compared pair or group.
For example, in the category of different school types (Table 4.13), at grade 9 level, the number of subjects (N), was determined by using the smallest number of subjects among the different school types [formerly DET, formerly model C and Private schools]. In this case, the Private school had 28 subjects in grade 9, while the formerly DET and model C schools had 132 and 50 subjects respectively, therefore 28 grade 9 subjects were randomly selected from the formerly DET and Model C school types. Such that, all the three groups compared, had the same number of subjects (28 each). The number of subjects compared in each category, were determined in the same way.

4.6.1 COMPARISON OF THE PERFORMANCE OF GIRLS AND BOYS

TABLE 4.9. COMPARISON OF THE PERFORMANCE OF GIRLS AND BOYS

(a) DESCRIPTIVES

Gender	N	$\bar{x}$	SD	S EM
Male	**57**	**33.25**	**12.128**	**1.606**
Female	**57**	**35.47**	**12.544**	**1.662**

KEY
N = Number of subjects
$\bar{x}$ = Average performance
SEM = Standard Error of Measurement.

(b) INDEPENDENT SAMPLE T-TEST

	Levene's test for equality of variance		t-test for equality of means			95% Confidence level	
Mark	**F**	**Sig (p)**	**t**	**df**	**Sig. (2-tailed)**	**$\bar{x}$ difference**	**Std error differ.**
Equal variance assumed	**0.218**	**0.642**	**-0.964**	**112**	**0.337**	**-2.228**	**2.311**
Equal variance not assumed			**-0.964**	**111.9**	**0.337**	**-2.228**	**2.311**

The results on table 4.9b show a t- value of – 0.964, with $p \leq 0.337$. This p-value is more than 0.05. This means that there is no significant difference in the mean performance of girls and boys on the developed test. The schools used in the study were all co-educational schools. The boys and girls compared were coming from the same classes. Hence it was assumed that boys and girls in the same class were subjected to the same conditions of teaching and learning. In other words, the other variables that could have affected the performance of the learners were constant in both groups. This result therefore suggests that the test is not gender biased.

4.6.2 COMPARISON OF THE PERFORMANCE OF LEARNERS FROM RURAL AND URBAN SCHOOLS

TABLE 4.10. COMPARISON OF THE PERFORMANCE OF LEARNERS FROM RURAL AND URBAN SCHOOLS

(a) DESCRIPTIVES

Location	**n**	**$\bar{x}$**	**SD**	**S EM**
Urban	**180**	**48.69**	**14.475**	**1.079**
Rural	**180**	**33.53**	**11.337**	**0.845**

KEY
N = Number of subjects
$\bar{x}$ = Average performance
SEM = Standard Error of Measurement.

(b) INDEPENDENT SAMPLE T-TEST

	Levene's test for equality of variance		t-test for equality of means			95% Confidence level	
Mark	**F**	**Sig (p)**	**t**	**df**	**Sig. (2-tailed)**	**$\bar{x}$ difference**	**Std error differ.**
Equal variance assumed	**13.900**	**0.000**	**11.063***	**358**	**0.000***	**15.161**	**1.370**
Equal variance not assumed			**11.063***	**338.56**	**0.000***	**15.161**	**1.370**

***The mean difference is significant at the 0.05 confidence level**

Table 4.10b shows the comparison between the performance of learners from urban and rural schools. A t-value of 11.063 was obtained, with $p \leq 0.000$. This p-value is less than 0.05, which means that the performance of the two groups is significantly different. The performance means of the groups of subjects involved [48.69 for urban schools, and 33.53 for rural schools] (Table 4.10a) shows quite a big difference.

On the surface, the conclusion from this result would be that, the test is biased against learners from rural schools. However, there are several factors that are likely to contribute to the poor performance of learners from rural schools. These factors include the following: first, most rural schools are not well equipped in terms of physical and laboratories facilities, which can negatively impact on the acquisition of science process skills.

Second, most rural schools lack teachers who are qualified to teach science. The country as a whole has few qualified science teachers (Zaaiman, 1998), and most are located in the cities, townships and urban areas. Lastly, most rural schools have very large science classes, in terms of teacher-pupil ratio. This makes the teaching and learning of science to be undertaken in ways that help teachers to cope with the large classes. And this is usually through chalk and talk transmission mode.

In summary, the conditions of teaching and learning in urban and rural schools are not the same. The significant difference observed in the performance of the two groups of learners may not therefore be attributed to the bias of the test instrument. A conclusive argument regarding the bias of the instrument against rural schools can only be reached if the two sets of schools being compared were subjected to similar teaching and learning conditions prior to the administration of the test.

The mean difference observed in the performance of rural and urban subjects may be an indication of the discrimination power and sensitivity of the developed test, in terms of its ability to identify learners who are more competent in integrated science process skills, and those who are less competent, presumably the urban and rural learners respectively.

4.6.3 COMPARISON OF THE PERFORMANCE OF WHITE AND BLACK LEARNERS.

TABLE 4.11 COMPARISON OF THE PERFORMANCE OF WHITE AND BLACK LEARNERS.

(a) DESCRIPTIVES

Race	n	$\bar{x}$	SD	S EM
White	**30**	**54.90**	**16.016**	**2.924**
Black	**30**	**54.93**	**13.821**	**2.523**

KEY
N = Number of subjects
$\bar{x}$ = Average performance
SEM = Standard Error of Measurement

(b) INDEPENDENT SAMPLE t-TEST

	Levene's test for equality of variance		t-test for equality of means			95% Confidence level	
Mark	**F**	**Sig (p)**	**t**	**df**	**Sig. (2-tailed)**	**$\bar{x}$ difference**	**Std error differ.**
Equal variance assumed	**2.173**	**0.136**	**-0.009**	**58**	**0.993**	**-0.033**	**3.862**
Equal variance not assumed			**-0.009**	**56.785**	**0.993**	**-0.033**	**3.862**

Table 4.11, compares the performance of white and black learners on the developed test. The statistics (Table 4.11b) show a t-value of – 0.009 (absolute value) with $p \leq 0.993$, which is more than 0.05. This means that the performance of white and black learners on the test was not significantly different.

The subjects were taken from the same classes in the same school, and each of the grades considered had both white and black learners, who presumably, were subjected to the same teaching and learning conditions. Thus the teaching and learning conditions for the two groups of learners were constant for both groups. The obtained result therefore suggests that the test was not biased against black or white learners.

4.6.4 COMPARISON OF THE PERFORMANCE OF LEARNERS ON THE DEVELOPED TEST AND ON TIPS.

One of the main arguments in this study was that, though the foreign developed tests of science process skills are valid and reliable when used for the target population, they are likely to disadvantage South African learners in a number of ways. The main disadvantage being that, the technical language and examples used in these tests are sometimes unfamiliar to the South African beginning science learners. As a result, learners may perform poorly, not because they are incompetent in the science process skills being tested, but because they are unable to relate in a meaningful way to the language and examples of the tests.

In this study, 30 subjects from each grade level were randomly selected from the school where the developed test and the TIPS [a foreign standardized test] (Dillashaw and Okey, 1980) were administered concurrently. Each of the 30 subjects in each grade level therefore had a pair of scores. One score from the developed test, and the other from the TIPS (Appendix X.). The mean scores from the two tests were compared according to the grade levels, as shown on table 4.12.

TABLE 4.12. COMPARISON OF THE PERFORMANCE OF LEARNERS ON THE DEVELOPED TEST AND ON TIPS.

(a) DESCRIPTIVES

Grade	Pair	N	$\bar{x}$ (%)	$\bar{x}$ difference	SD	SEM	Correlation
9	DT & TIPS	30	64.63 50.80	13.833	9.565 10.084	1.746 1.841	0.503
10	DT & TIPS	30	63.53 51.57	11.967	12.958 13.890	2.366 2.536	0.568
11	DT & TIPS	30	71.93 60.53	11.400	8.902 10.750	1.625 1.963	0.599

KEY
N = Number of subjects
$\bar{x}$ = Average performance
SEM = Standard Error of Measurement

(b) PAIRED SAMPLES T-TEST

		Paired differences					
Grade	Pair	$\bar{x}$	SD	SEM	t	df	Sig (2-tailed)
9	DT & TIPS	13.833	9.805	1.790	7.727*	29	0.000*
10	DT & TIPS	11.967	12.502	2.283	5.243*	29	0.000*
11	DT & TIPS	11.400	8.958	1.636	6.970*	29	0.000*

***The mean difference is significant at the 0.05 confidence level**

Table 4.12a, shows the statistics for the comparison of learners' performance on a standard test (TIPS), and the developed test, using the paired samples t-test. The data show t-values of; 7.727 for grade 9, 5.243 for grade 10, and 6.970 for grade 11, and they all have $p \leq 0.000$, which is less than 0.05. This suggests that, the difference in the performance of learners on the developed test and TIPS was significantly different in all grades. In each grade, the performance of the learners on the developed test (DT) was higher than that on the standard test (TIPS). This is clearly evident from the mean performance of learners in all grades (for grade 9, DT = 64.63 **:** 50.80 = TIPS; for grade 10, DT = 63.53 **:** 51.57 = TIPS; and for grade 11, DT = 71.93 **:** 60.53 = TIPS).

The two tests assess the same science process skills, and are referenced to the same set of objectives. They also have the same multiple-choice format. The difference between the two tests, which might have caused the observed discrepancy is that, the developed test does not use foreign examples and technical terms, while the standard test does. These results aside from testing concurrent validity also support the argument that the foreign developed tests place local learners at a disadvantage.

4.6.5 COMPARISON OF THE PERFORMANCE OF LEARNERS FROM DIFFERENT SCHOOL TYPES

Table 4.13 below, compares the performance of learners from different school types, that is; formerly DET, formerly model C, and private schools. The analysis of the results on table 4.13 were done according to grades, because the different grade levels of the different school types showed different performance results.

For grade 9 learners, the results show an F-value of 19.017 with $p \leq 0.000$ (Table 4.13b). This p-value is less than 0.05, suggesting that there was a significant difference in the performance of grade 9 learners coming from different school types. The multiple comparisons (Table 4.13c) show that learners from former model C schools performed much better than those from former DET and private schools, and there was no significant difference between the performance of learners from private and formerly DET schools.

TABLE 4.13. COMPARISON OF THE PERFORMANCE OF LEARNERS FROM DIFFERENT SCHOOL TYPES

(a) DESCRPTIVES

Grade	Type	N	$\bar{x}$	SD	SEM
9	Model C	28	52.64	11.525	2.178
	DET	28	35.50	9.191	1.737
	Private	28	38.71	12.226	2.310
	Total	**84**	**42.29**	**13.242**	**1.445**
10	Model C	30	51.90	13.525	2.469
	DET	30	46.33	12.691	2.317
	Private	30	53.13	10.513	1.919
	Total	**90**	**50.46**	**12.528**	**1.321**
11	Model C	25	55.00	18.755	3.751
	DET	25	52.16	10.015	2.003
	Private	25	74.44	10.034	2.007
	Total	**75**	**60.53**	**16.692**	**1.927**

KEY
N = Number of subjects
$\bar{x}$ = Average performance
SEM = Standard Error of Measurement

(b) ONE-WAY ANOVA

Grade		SS	df	M S	F	Level of Sig.
9	**Between groups**	**4650.000**	**2**	**2325.00**	**19.017***	**0.000***
	Within groups	**9903.143**	**81**	**122.261**		
	Total	**14553.143**	**83**			
10	**Between groups**	**787.489**	**2**	**393.744**	**2.599**	**0.080**
	Within groups	**13180.833**	**87**	**151.504**		
	Total	**13968.322**	**89**			
11	**Between groups**	**7353.147**	**2**	**3676.573**	**19.955***	**0.000***
	Within groups	**13265.520**	**72**	**184.243**		
	Total	**20618.667**	**74**			

***The mean difference is significant at the 0.05 confidence level**

(c) MULTIPLE COMPARISONS

Grade	(I) Type	(J) Type	$\bar{X}$ difference (I – J)	Std error	Sig	Lower limit	Upper limit
9	Model C	DET	17.143*	2.955	0.000	9.92	24.37
		Private	13.929*	2.955	0.000	6.70	21.15
	DET	Model C	-17.143*	2.955	0.000	-24.37	-9.92
		Private	-3.214	2.955	0.840	-10.44	4.01
	Private	Model c	-13.929*	2.955	0.000	-21.15	-6.70
		DET	3.214	2.955	0.840	-4.01	10.44
11	Model C	DET	2.840	3.839	1.000	-6.57	12.25
		Private	-19.440*	3.839	0.000	-28.85	-10.03
	DET	Model C	-2.840	3.839	1.000	-12.25	6.57
		Private	-22.280*	3.839	0.000	-31.69	-12.87
	Private	Model c	19.440*	3.839	0.000	10.03	28.85
		DET	22.280*	3.839	0.000	12.87	31.69

***The mean difference is significant at the 0.05 confidence level**

For grade 10 learners, the F-value of 2.599 with $p \leq 0.080$ was obtained (Table 4.13 b). This p-value is higher than 0.05. This means that there was no significant difference in the performance of grade 10 learners from the different school types. As a result, multiple comparison of the performance of grade 10 learners from the different school types was not necessary. This result is interesting, given the varied teaching and learning conditions in these schools. One would have expected significant difference in the performance of these learners, as the case is in other grades. The result may however be explained in terms of the learner exodus that happens at grade 10 level. This stage marks the transition from the General Education and Training (GET) band to the Further Education and Training (FET) band.

During this transition, several learners move from one type of school to another. This makes the grade 10 classes from the different school types constitute a mixed ability group of learners (coming from different school types), getting adjusted to their new environment. Hence, with the mixed ability groups in the different school types, the mean performance of the grade 10 learners is likely to be uniform. In other words, it is unlikely that there may be a significant difference in the mean performance of grade 10 learners from the different school types.

The statistics for grade 11 learners show an F value of 19.955 with $p \leq 0.000$ (Table 4.13b). In this case, the differences observed in the mean performance of learners from different school types were significant. The multiple comparison of the performance of learners from different school types (Table 13c), show that there was no significant difference between the performance of formerly model C, and formerly DET learners, but there was a significant difference in performance of learners from formerly model C schools and private schools. There was also a significant difference in the performance of learners from formerly DET and private schools (Table 4.13c). The mean performance of learners as evident from table 4.13a, shows that the learners from private schools performed better than those from formerly model C and DET schools.

The overall result of this analysis implies that the developed test is sensitive to some school types.

4.6.6 COMPARISON OF THE PERFORMANCE OF LEARNERS FROM DIFFERENT GRADE LEVELS

TABLE 4.14. COMPARISON OF THE PERFORMANCE OF LEARNERS FROM DIFFERENT GRADE LEVELS

(a) DESCRIPTIVES

Grade	N	$\bar{x}$	SD
9	**120**	**38.52**	**12.9**
10	**120**	**41.73**	**13.4**
11	**120**	**50.48**	**15.8**

KEY
N = Number of subjects
$\bar{x}$ = Average performance
SD = Standard Deviation

(b) ONE –WAY ANOVA

	SS	df	MS	F	Level of Sig.
Between groups	**9204.422**	**2**	**4602.211**	**20.412***	**0.000***
Within groups	**80489.400**	**357**	**225.461**		
Total	**89693.822**	**359**			

***The mean difference is significant at the 0.05 confidence level**

(c) MULTIPLE COMPARISONS

(I) grade	(J) grade	$\bar{x}$ difference (I – J)	Std error	Sig	Lower limit	Upper limit
9	**10** **11**	**-3.217** **-11.967***	**1.938** **1.938**	**0.294** **0.000***	**-7.88** **-16.63**	**1.45** **-7.30**
10	**9** **11**	**3.217** **-8.750***	**1.938** **1.938**	**0.294** **0.000***	**-1.45** **-13.41**	**7.88** **-4.09**
11	**9** **10**	**11.967*** **8.750***	**1.938** **1.938**	**0.000*** **0.000***	**7.30** **4.09**	**16.63** **13.41**

***The mean difference is significant at the 0.05 confidence level**

Table 4.14 compares the performance of learners from the different grade levels, to establish whether the differences observed in their performance were significant. The one-way ANOVA results show an F-value of 20.412, with $p \leq 0.000$ (Table 4.14b), which is less that 0.05 (confidence level). These results show that there was a significant difference in the performance of learners in the three grades.

A multiple comparison of the performance of learners from different grades shows that there was a significant difference in the performance of grade 9 and grade 11 learners, as well as between grade 10 and grade 11 learners. There was no significant difference in the performance of grade 9 and grade 10 learners, as indicated in table 4.14c. This result can also be seen from an inspection of the grade means (Table 4.14a). The high mean performance of grade 11 learners implies that the grade 11 learners found the test to be easier than the grade 9 and 10 learners. This is confirmed by the overall difficulty index for grade 11 learners, which is much higher than that of the grade 9 and 10 learners (Table 4.7).

The high performance of grade 11 learners compared to the lower grades is expected, because learners in higher grades are likely to have had more experience with activities involving process skills and the subject content, than those in lower grades. This result suggests that the test is sensitive, and it has a good discrimination power.

A summary of the results from the comparison of the different groups, in the different categories are displayed in table 4.15, below. The table shows the differences in the means of the compared groups, the p values, and the significance of the mean differences.

TABLE 4.15 SUMMARY OF THE COMPARISON OF THE PERFORMANCE OF DIFFERENT GROUPS OF LEARNERS. [At 0.05 (95%) confidence level].

CATEGORY	GROUPS		$\bar{x}$ DIFFERENCE	P- VALUE	COMMENT
GENDER	GIRLS V BOYS		2.228	0.337	Not significant
LOCATION	RURAL V URBAN		15.161*	0.000	Significant
RACE	BLACK V WHITE		0.033	0.993	Not significant
TYPE OF TEST	DEV TEST V TIPS	9	13.833*	0.000*	Significant
		10	11.967*	0.000*	Significant
		11	11.400*	0.000*	Significant
GRADES	9 VERSUS 10		3.217	0.294	Not significant
	10 VERSUS 11		8.750*	0.000*	Significant
	11 VERSUS 9		11.967*	0.000*	Significant
SCHOOL TYPE	PRIVATE V DET	9	3.214	0.840	Not significant
		10	6.80	0.080	Not significant
		11	22.280*	0.000*	Significant
	DET V MODEL C	9	17.143*	0.000*	Significant
		10	5.570	0.080	Not significant
		11	2.840	1.000	Not significant
	MODEL C V PRIVATE	9	13.929*	0.000*	Significant
		10	1.23	0.080	Not significant
		11	19.440*	0.000*	Significant

***The mean difference is significant at the 0.05 confidence level**

CHAPTER 5
CONCLUSIONS

This chapter presents a summary of the results and the conclusions made from them, as well as their implications for the educational system. The chapter further highlights the recommendations based on the findings, the limitations of the study, and areas for further research.

The main aim of this study was to develop and validate, a reliable and convenient test, for measuring integrated science process skills competence, effectively and objectively in schools. The science process skills tested for were; identifying and controlling variables, stating hypotheses, designing investigations, graphing and interpreting data, and operational definitions.

In order to achieve the above stated aim, the paper and pencil group-testing format was used in this study. Thirty (30) multiple-choice items (see Appendix I), referenced to nine (9) specific objectives (Table 3.3), were developed and validated, after a series of item analysis, reviews and modifications. The test items were constructed in a way that tried to eliminate bias towards different groups of learners. The items were administered to seven hundred and sixty nine (769) grade 9, 10 and 11 learners from the Capricorn district of the Limpopo province, in the main study.

5.1 SUMMARY OF RESULTS, AND CONCLUSIONS

The results of the study show that the test characteristics of the developed instrument fall within the acceptable range of values as shown in table 5.1 below. This suggests that the developed instrument is valid and reliable enough, to be used to measure learners' competence in the stated science process skills, in the further education and training band.

TABLE 5.1. SUMMARY OF THE TEST CHARACTERISTICS OF THE DEVELOPED INSTRUMENT.

Test characteristic	Overall	Acceptable values
Discrimination index	**0.403201**	≥ **0.3**
Index of difficulty	**0.401853**	**0.4 – 0.6**
Content validity	**0.97846**	≥ **0.7**
Concurrent validity /alternative form reliability	**0.56**	≥ **0.7**
Reliability	**0.81**	≥ **0.7**
Standard Error of Measurement (SEM)	**7.0671**	**Not specified**
Readability level	**70.2902**	**60 - 70**
Reading grade level	**Grade 8**	**Grades 9, 10, 11**

The first research question, which sought to determine whether the developed test could be shown to be a valid and reliable means of measuring integrated science process skills competence in schools, is therefore satisfied. It should be noted however that the concurrent validity [whose value is below the accepted range of values for validity] (Table 5.1) may not be considered in this conclusion, for reasons earlier advanced (section 4.6.4).

The paper and pencil group testing format does not require expensive resources, and it can easily be administered to large groups of learners at the same time, hence it may be concluded that the test is cost effective and convenient.

The second research question concerned the fairness of the test, that is, if the developed test instrument could be shown to be location, school type, race, and gender neutral. The results from the comparison of the performance of different groups of learners show that there was no significant difference between the performance of white and black learners, and between boys and girls (Table 4.15). This result suggests that the test instrument is not race or gender biased.

The results from table 4.15 also show that there was a significant difference in the mean performance of learners from rural and urban schools. As discussed in section 4.6.2, this result may not be interpreted as an indication of test bias against rural schools, due to the variability of the teaching and learning conditions prevalent in the two systems. The differences in the mean performance of learners from different school types were significant in some cases and insignificant in others as shown on table 4.15, due to the varied nature of the schools involved, in terms of teaching and learning conditions, as discussed in section 4.6.5. These results show that the developed test is sensitive and discriminatory in regard to the acquisition of science process skills.

The significant difference observed among the different grade levels (Table 4.15) may be considered as an indication that the test has a good discrimination power, since it can discriminate between those who are likely to be more competent in science process skills (grade 11 learners) and those who are likely to be less competent in the skills (grade 9 learners).

It may be concluded therefore that the second research question was also satisfied, in that, the test was shown to be gender and racial neutral (sections 4.6.1 and 4.6.3), and that it is sensitive and can discriminate well among learners who have acquired the measured science process skills and those who have not (sections 4.6.2 and 4.6.5).

The results further show that there was a significant difference between the performance of learners on the developed test and a standard test (TIPS). The performance of learners was higher on the developed test than on the standard test used (TIPS), in all grades (Table 4.12a). These results are in agreement with the argument that the foreign developed tests may not always be suitable for South African learners.

5.2 EDUCATIONAL IMPLICATIONS OF RESULTS

Based on the results of this study, The educational implications of the study may be summarized as follows;

- The aim of this study was to develop a test instrument that could be directly used by science educators to assess their learners' competence in integrated science process skills. This study contributed to education under the research and development category, which is described by Gay (1987) as research that is directed at the development of effective products that can be used in schools.

- The test instrument was constructed in such a way that it is user friendly within the South African context. The study may therefore be considered as an improvement on similar instruments that are currently presenting challenges to the South African users.

- The instrument developed from this study may be used to collect information about how well learners are performing in the acquisition of integrated science process skills, and thus contribute to the description of educational phenomenon.

- As stated in earlier (section 1.1), Science education in South Africa is characterized by practical constraints that make the traditional assessment of science process skills through practical work rather cumbersome and unfulfilled in some instances. These constraints which include lack of resources, over crowded large classes, ill-equipped laboratories, unqualified or under qualified science educators, etc, may be abated or overcome through the use of the instrument developed from this study, as an alternative assessment tool of higher order thinking in science.

- The language used in assessment tests has been known to influence the learners' performance (Kamper, Mahlobo and Lemmer, 2003). The discussion of the results of this study alluded to the fact that language facility and familiarity with technical words may affect learners' demonstration of competence in science process skills. Care must therefore be taken to ensure that language or lack of familiarity do not become stumbling blocks when assessing learners' competence in any area of study. The study also provides empirical support (concurrent validity) that the use of foreign terminology and technical terms in process skills tests is likely to disadvantage some learners who may perform poorly because of a lack of comprehension of terms.

5.3 RECOMMENDATIONS

- The developed instrument could be readily adapted to local use to monitor the acquisition of science process skills by the learners. The results of which could feedback on the effectiveness of the new science curriculum.

- The developed instrument could be used by researchers in various ways. For instance, researchers who need a valid and reliable instrument to work with, may use the test to; identify the process skills inherent in certain curricula material, determine the level of acquisition of science process skills in a particular unit; establish science process skills competence by science teachers, or to compare the efficacy of different teaching methods in imparting science process skills to learners.

- Researchers could also use the procedure used to develop the test instrument as a model for the development and validation of other similar assessment instruments.

- The paper and pencil test is a convenient, efficient, and cost effective tool, which may be used by educators for classroom assessment of learners' competence in integrated science process skills. It could be used for baseline, diagnostic, continuous, or formative assessment purposes, especially by those teaching poorly resourced large classes.

- Furthermore, being a multiple-choice test, the developed test could be administered anywhere at any time by anyone with or without expertise in the field of science process skills. Moreover, marking of the test will be consistent, reliable and greatly simplified.

- Lastly, learners and their teachers could use the developed instrument to get prompt feedback on their competence in science process skills, so that they are able to identify areas where they may need remediation.

5.4 LIMITATIONS OF THE STUDY

The study has certain limitations that should be taken into consideration when interpreting the results. The limitations pertain to the following;

- The test instrument is meant for learners from the further education and training bands. These bands involve grades 10, 11 and 12 learners. The study however only involved grades 9, 10 and 11. The exclusion of grade12 learners from the study may not present a complete picture of the performance of the test instrument in the designated band.

- A criticism of multiple-choice questions is that candidates cannot justify their choices. This may be avoided by making a provision for candidates to explain or justify their choices. This approach eliminates the possibility of guessing, which is prevalent in multiple-choice type of tests.

- The use of a paper and pencil test to assess practical skills has been criticized by several researchers who advocate for practical manipulation of apparatus and physical demonstration of practical skills. This presents a limitation in the sense that the instrument developed in this study does to accommodate these requirements.

- The developed test was compared with (TIPS) to determine its external validity. TIPS, however, has some constraints which could have led to the learners' poor performance on it. Comparison of performance of learners on the developed test with their performance in any other alternative locally developed assessment instrument could perhaps have been a better criterion to use in determining the external or concurrent validity of the developed test instrument.

5.5 AREAS FOR FURTHER RESEARCH

The results from this study present several further research opportunities, which include the following:

- The instrument maybe used to determine competence of teachers in integrated science process skills.
- An instrument, which tests competence in primary science process skills may be developed and validated, based on the format and methodology used in this study.
- The instrument may be used to assess learners' competence in integrated science process skills nationally, to determine the effectiveness of the new curriculum in imparting science process skills to learners.

REFERENCE

Adkins, D.C., (1974). *Test construction: Development and interpretation of achievement tests.* Columbus, Ohio: Charles E. Merrill publishing Co.

Arnott, A., Kubeka, Z., Rice, M., & Hall, G. (1997). Mathematics and Science Teachers: Demand, utilization, supply and training in South Africa. *Edu-source* 97/01. Johannesburg: The Education Foundation.

Atkinson, E. (2000). In Defense of Ideas, or Why "What works" is not enough. *British Journal of the Sociology of Education*, 2 (3), 317-330.

Baird, W.E., & Borich, G.D. (1985). Validity Considerations for the Study of Formal Reasoning Ability and Integrated Science Process Skills. ERIC No: ED254428, Paper presented at the Annual Meeting of the National Association for Research in Science Teaching, (58th, French Lick Springs, IN, April 15-18, 1985).

Basterra, M.R (1999). Using standardized tests to make high stake decisions on English-Language learners: dilemmas and critical issues. *Equity Review*. Spring 1999. Retrieved on 26th July, 2004, from: http://www.maec.org/ereview1.html

Bates, G.R. (2002). The impact of Educational Research: Alternative Methodologies and Conclusions. *Research papers in Education*, (Submitted but not published). Contact: rbates@deakin.edu.au. Deakin University, Australia.

Berk, R.A. (Ed). (1982). *Handbook of methods for detecting test bias*. Batimore, M.D. The Johns Hopkins University Press.

Berry, A., Mulhall, P., Loughran, J.J., & Gunstone, R.F. (1999). Helping students learn from Laboratory work. *Australian Science Teachers' Journal*, 45(1), 27-31.

Bloom, B. S., Englehart, M. D., Furst, E. J., & Krathwohl, D. R. (1956). *Taxonomy of Educational objectives: The Classification of Educational goals.* Handbook 1: Cognitive domain. White Plains, New York. Taxonomy of Educational objectives: The Classification of Educational goals. Handbook 1: Cognitive domain. White Plains, New York: Long man, LB17. T3. Long man, LB17. T3.

Bredderman, T. (1983). Effects of Activity Based elementary Science on Student Outcomes: A Qualitative Synthesis. *Review of Educational Research*. 53 (4), 499-518.

Brescia, W., & Fortune J.C. (1988). Standardized Testing of American Indian Students. *Las Cruces,* NM 88003-0001 (505) 646-26-23: ERIC CRES.

Brown, J.D. (1996). *Testing in language programmes*. Upper Saddle River. N.J: Prentice Hall Regents.

Brown, J.D. (2000 Autumn). What is construct validity? TALT, *Testing and evaluation SIG News letter*, 4 (2), 7-10.

Burns, J.C., Okey, J.R., & Wise, K.C. (1985). Development of an Integrated Process Skills Test: TIPS II. *Journal of Research in Science Teaching*. 22 (2). 169-177.

Carneson, J., Delpierre, G. and Masters, K. (2003). *Designing and managing multiple-choice questions*. Australia. Southrock Corporative Limited.

Childs, R.K. (1990). *Gender Bias and Fairness*. Eric Digest, Ed. 328610. Washington D.C.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. NewYork: Holt, Rinehart and Winston.

Department of education, 2002. *Revised National Curriculum Statement, Grades R – 9, Schools Policy, Natural Sciences*. Pretoria, South Africa: FormeSet Printers Cape.

Dietz, M. A., & George, K. D. (1970), A Test to Measure Problem Solving Skills in Science of Children in grades one, two and three, *Journal of Research in Science Education,* 7 (4), 341 – 351.

Dillashaw, F.G., & Okey, J.R. (1980). Test of Integrated Science Process Skills for Secondary Students. *Science Education*, 64, 601 – 608.

Fair Test Examiner. (1997). SAT Math gender gap: Causes and consequences. 342 Broadway, Cambridge MA. 02139. Retrieved on 22 July 2005, from, http://www.fairtest.org/examarts/winter97/gender.htm

Flier, H., Thijs, G.D., & Zaaiman, H. (2003). Selecting Students for a South African Mathematics and Science Foundation Program: The Effectiveness and Fairness of School-leaving Examinations and Aptitude Tests. *International Journal of Educational Development*, 23, 399-409

Froit, F.E. (1976). *Curriculum experiences and movement from concrete to operational thought. In Research, Teaching and Learning with the Piaget model*. Norman, U.S.A: University of Oklahoma Press.

Gall, M.D., Borg, W.R., & Gall, J.P. (1996). *Educational research: An Introduction*. New York, U.S.A: Longman publishers.

Gay, L.R. (1987). *Educational Research: Competencies for Analysis and Application*. 3rd ed. U.S.A: Merrill publishing Co.

Gott, R. and Duggan, S. (1996) Practical work: Its role in the understanding of the evidence in Science. *International Journal of Science Education, 18(7), 791-806.*

Hambleton, R., & Rodgers, J. (1995). Developing an Item Bias Review Form: [Electronic Journal] *Practical Assessment, Research and Evaluation*, 4 (6), Retrieved on 2nd August, 2004, from, http://pareonline.net/getvn.asp?v=4&n+6

Harlen, W. (1999). Purposes and Procedures for Assessing Science Process Skills. *Assessment in Education*, 6 (1) 129-135.

Higgins, E. and Tatham, L. (2003). Exploring the potential for multiple-choice questions in assessment. *Assessment 2 (4.shtlm) ISSN 1477-1241. Retrieved on 3rd February 200,. from http://www.Itu.mmu.ac.uk/Itia/issue4/higginstatham.shtml*

Hinkle, W.J. (1998). *Applied Statistics for the Behavioral Sciences*. 4th ed, Boston: Houghton Mifflin Company

Howe, K. (1995). Validity, Bias and Justice in Educational Testing: The Limits of Consequentialist Conception. *Philosophy of Education*. Retrieved on 14th June, 2004, from: Http://www.ed.uiuc/EPS/PES~yearbook/95_docs.howe.html

Howie, S.J. (2001). *Third International Mathematics and Science Study*. (Paper presented at the 1st National Research Coordinators Meeting, 25-28 February, 2001. Hamburg, Germany). From Education and Training - A report on a HSRC Trip. HSRC Library, shelf no. 1882.

Howie, S.J., & Plomp, T. (2002). Mathematical literacy of school learning pupils in South Africa. *International Journal of Educational development,* 22, 603- 615.

HSRC. (2005a). Survey gives hard facts about the lives of educators. Human Sciences Research Council Review. 3 (2), July 2005.

HSRC. (2005b). Research highlights. Human Sciences Research Council Annual report 2004/2005. Retrieved on 31st October, 2005, from: http:www.hsrc.ac.za/about/annual Rep…/researchHighlights.htm

HSRC. (1997). *School Needs Analysis*. Human Sciences Research Council. Pretoria. Retrieved on 13th January, 2004, from: http://www.hsrcpublishers.co.za/

Kamper, G.D., Mahlobo, E.B., & Lemmer, E.M. (2003). The relationship between standardized test performance and language learning strategies in English second Language: A case study. *Journal for Language Teaching* (An on-line journal), 37 (2). Retrieved on 12 /04/2005, from http://www.language.tut.ac.za/saalt/jour376-2.htm

Klare, G. (1976). A Second Look at the Validity of Readability Formulas. *Journal of Reading Behaviour*, 8, 129-152.

Lavinghousez, W.E., Jr. (1973). The analysis of the Biology Readiness Scale (BRS), as a measure of inquiry skills required in BSCS Biology. College of education, University of central Florida. February 25, 1973.

Magagula, C.M., & Mazibuko, E.Z. (2004). Indegenization of African Formal Educational Systems. *The African symposium* (An on-line journal). 4 (2). Retrieved on 13/4/2005, from http://www2.ncsu.edu/ncsu/aern/inafriedu.htm

McLeod, R.J., Berkheimer, G.G., Fyffe, D.W., & Robinson, R.W. (1975). The Development of Criterion-validated Test items for four Integrated Science Processes. *Journal of Research in Science Teaching*, 12, 415-421.

Messick, S. (1988). The Once and Future Issues of Validity. Assessing the meaning and consequences of measurement. In H. Wainer and H.I. Braun (Eds) *Test Validity*, pp33-45. Hillsdale, NJ. Lawrence Erlbaum Associates.

Millar, R., Lubben, F., Gott, R. and Duggan, S. (1994). Investigating in the school laboratory: Conceptual and Procedural Knowledge and their Influence on Performance. *Research Papers in Education, 9(2), 207-248.*

Millar, R. and Driver, R. (1987). Beyond Processes. *Studies in Science Education, 14, 33-62.*

Molitor, L.L., & George, K.D. (1976). Development of a Test of Science Process Skills. *Journal of Research in Science Teaching*, 13(5), 405 – 412.

Mozube, B.O. (1987). *Development and Validation of Science Skills Test for Secondary Schools*. Unpublished Masters dissertation, 1987, University of Ibadan, Nigeria.

Muwanga-Zake, J.W.F. (2001a). Is Science Education in a crisis? Some of the problems in South Africa. *Science in Africa* (Science on-line Magazine), Issue 2. Retrieved on 13/04/2005, from http://www.scienceinafrica.co.za/scicrisis.htm

Muwanga-Zake, J.W.F. (2001b). Experiences of the Power of Discourse in Research: A need for transformation in Educational Research in Africa. *Educational Research in Africa* (on-line Science magazine), 1 (4). Retrieved on 21/04/2005, from http://www2.ncsu.edu/ncsu/aern/muwangap.html

National Academy of Sciences. (1995). *U.S.A National Science Education Standards*. Retrieved on 10th September, 2003, from: http://www.nap.edu.nap.online/nses.

National Department of Education. (1995). White Paper on Education and Training. Notice 196 of 1995, WPJ/1995. Cape town.

National Department of Education. (1996). Education White Paper-2; The organization, governance, and funding of schools. Notice 130 of 1996. Pretoria.

Nitko, A.J. (1996). *Educational assessment of students*. 2nd ed. New Jersey, U.S.A. Prentice-Hall.

Novak, J.D. & Govin, D.B. (1984). *Learning how to learn*. Cambridge: Cambridge University Press.

Novak, J.D., Herman, J.L. & Gearhart, M. (1996). Establishing Validity for Performance Based Assessments: An illustration for Student Writings. *Journal of Educational Research*, 89 (4), 220 – 232.

Okrah, K.A. (2004). African Educational Reforms in the era of Globalization: Conflict or Harmont? *The African Symposium* (An on-line journal), 4(4). Retrieved on 13/04/2005, from http://www2.ncsu.edu/ncsu/aern/okrahdec04.htm

Okrah, K.A. (2003). Language, Education and culture – The Dilemma of Ghanaian Schools. *The African Symposium* (An on-line journal). Retrieved on 13/04/2005, from http://www2.ncsu.edu/ncsu/aern/inkralang.html

Okrah, K.A. (2002). Academic colonization and Africa's underdevelopment. *The African Symposium* (An on-line journal). Retrieved on 13/04/2005, from http://www2.ncsu.edu/ncsu/aern/okrapgy.html

Onwu, G.O.M, and Mogari, D (2004). Professional Development for Outcomes based Education Curriculum imlementation: The case of UNIVEMALASHI, South Africa. *Journal of Education for teaching.* 30 (2), 161-177.

Onwu, G.O.M (1999). Inquiring into the concept of large classes: Emerging topologies in an African context. In Savage, M. & Naidoo, P. (Eds.) *Using the local resource base to teach Science and Technology. Lessons from Africa*. AFCLIST. October, 1999.

Onwu, G.O.M. (1998) Teaching large classes. In Savage, M. & Naidoo, P. (Eds.) *African Science and Technology Education. Into the new millenium: Practice. Policy and Priorities.* Juta: Cape Town.

Onwu, G.O.M., & Mozube, B. (1992). Development and Validation of a Science Process Skills Test for Secondary Science Students. *Journal of Science Teachers' Association of Nigeria,* 27 (2), 37-43.

Osborne, R., & Freyberg, P. (1985). *Learning in Science: The implications of children's science*. Auckland, London: Heinemann publishers.

Ostlund, K. (1998). What Research Says about Science Process Skills. *Electronic Journal of Science Education*, 2 (4), ISSN 1087-3430. Retrieved on 17th February from: http://unr.edu/homepage/jcannon/ejse/ejsev2n4

Padilla, M.J. (1990). *Research Matters – To the Science Teacher*. No. 9004. March 1, 1990. University of Georgia. Athens. G.A.

Padilla, M.J., Mckenzie, D.L., & Shaw, E.L. (1986). An Examination of the line graphing skills Ability of students in grades seven through twelve. *School Science and Mathematics*, 86 (1), 20 –29.

Padilla, M.J., *et al*. (1981). *The Relationship Between Science Process Skills and Formal Thinking Abilities.* ERIC NO: ED201488, (Paper presented at the Annual Meeting of the National Association for Research in Science Teaching, (54th, Grossinger's in the Catskills, Ellenville, NY, April 5-8,1981).

Pollitt, A., & Ahmed, A. (2001). *Science or Reading?: How do students think, when answering TIMSS questions?* (A paper presented to the International Association for Educational Assessment). Brazil, May 2001.

Pollitt, A., Marriott, C., & Ahmed, A. (2000). *Language, Contextual and Cultural Constraints on Examination Performance*. A paper presented to the International Association for Educational Assessment, in Jerusalem, Israel, May 2000.

Rezba, R.J., Sparague, C.S., Fiel, R.L., Funk, H.J., Okey, J.R., & Jaus, H.H. (1995). Learning and Assessing Science Processes. (3rd Ed). Dubuque. Kendall/Hunt Publishing Company.

Ritter, M.J., Boone J.W., & Rubba, P.A. (2001). Development of an Instrument to Assess Perspective Elementary Teachers' Self-efficacy Beliefs about Equitable Teaching and Learning. *Journal of Science Teacher Education*,12 (3), 175 – 198.

Rosser, P. (1989). *The SAT gender gap: Identifying causes*. Washington, DC: Center for Women Policy Studies. ERIC Document Reproduction Service No. ED 311 087

Rudner, L. M. (1994). Questions to Ask when Evaluating Tests. *Electronic Journal of Practical Assessment, Research and Evaluation*, 4 (2). Retrieved August 2, 2004 from: http://pareonline.net/getvn.asp?v=4&n=2

Shann, M.H. (1977). Evaluation of Interdisciplinary Problem Solving Curriculum in elementary Science and Mathematics. *Science Education*, 61, 491-502

Simon, M.S., & Zimmerman, J.M. (1990). Science and Writing. *Science and Children*, 18 (3), 7-8.

Stephens, S. (2000). All about readability. Retrieved March 8, 2005, from: http://www.plainlanguagenetwork.org/stephens/readability.html

Tannenbaum, R.S. (1971). Development of the Test of Science Processes. *Journal of Research in Science Teaching*, 8 (2), 123-136.

The American Association for the Advancement of Science. (1998). *Blue prints for reform: Science Mathematics and Technology Education.* New York: Oxford University press.

Thomas, M., & Albee, J. (1998). *Higher order thinking strategies for the classroom*. (Paper presented at Mid-West Regional- ACSI, convention) Kansas city, October 1998.

Tipps, S. (1982). *Formal Operational Thinking of gifted students in grades 5, 6, 7, and 8*. (Paper presented at the annual meeting of the National Association for Research in Science Teaching) Lake Geneva, WI

Tobin, K.G., & Capie, W. (1982). Development and Validation of a Group Test of Integrated Science Process Skills," *Journal of Research in Science Teaching*, 19 (2), 133 – 141.

Trochium, W.M.K. (1999). *Research Methods: Knowledge Base*. 2nd Ed. Retrieved on 22nd September, 2003, from: File://C\M.Sc. Ed\Med. on line\VALIDI.HTM

Van de Vijver, F.J.R & Poortinga, Y.H. (1992). Testing in culturally heterogeneous populations: When are cultural loadings undesirable? *European Journal of Psychological Assessment*, 8, pp17-24.

Van de Vijver, F.J.R. & Hambleton, R.K. (1996). Translating Tests: Some practical guidelines. *European Psychologist*, 9, 147 – 157.

Walbesser, H.H. (1965). An evaluation model and its application. In the America Association for the Advancement of Science. *AAAS Miscellaneous publication* No. 65-9, Washington D.C.

Wiederhold, C. (1997). *The Q-Matrix/Cooperative learning and higher level thinking.* San Clemente, CA: Kagan Cooperative learning.

Wolming, S. (1998). Validity: A modern Approach to a Traditional Concept. *Pedogogisk Forskning: Sverige*, 3 (2). 81-103. Slotokholm. ISSN 1401- 6788.

Womer, F.B. (1968). *Basic concepts in testing*. Boston: Houghton Mifflin Co.

Zaaiman, H. (1998). *Selecting students for mathematics and Science: The challenge facing higher Education in South Africa*. South Africa: HSRC publishers. ISBN 0-7969-1892-9.

Zieky, M. (2002-Winter). Ensuring the Fairness of Licensing Tests. *Educational Testing Service*. From CLEAR, *Exam Review*, 12 (1), 20-26. http://www.clearing.org/cer.htm

APPENDICES

APPENDIX I.

THE TEST INSTRUMENT

TEST OF INTEGRATED SCIENCE PROCESS SKILLS

DURATION: 50 minutes

INSTRUCTIONS:

1. **VERY IMPORTANT!!!!!!!!!!!!!**

   DO NOT WRITE ANYTHING ON THE QUESTION PAPER.

2. **ANSWER ALL THE QUESTIONS ON THE ANSWER GRID PROVIDED, BY PUTTING A CROSS [X] ON THE LETTER OF YOUR CHOICE.**

3. **PLEASE DO NOT GIVE MORE THAN ONE ANSWER PER QUESTION.**

1. A learner wanted to know whether an increase in the amount of vitamins given to children results in increased growth.

   How can the learner measure how fast the children will grow?

   A. By counting the number of words the children can say at a given age.
   B. By weighing the amount of vitamins given to the children.
   C. By measuring the movements of the children.
   D. By weighing the children every week.

2. Nomsa wanted to know which of the three types of soil (clay, sandy and loamy), would be best for growing beans. She planted bean seedlings in three pots of the same size, but having different soil types. The pots were placed near a sunny window after pouring the same amount of water in them. The bean plants were examined at the end of ten days. Differences in their growth were recorded.

   Which factor do you think made a difference in the growth rates of the bean seedlings?

   A. The amount of sunlight available.
   B. The type of soil used.
   C. The temperature of the surroundings.
   D. The amount of chlorophyll present.

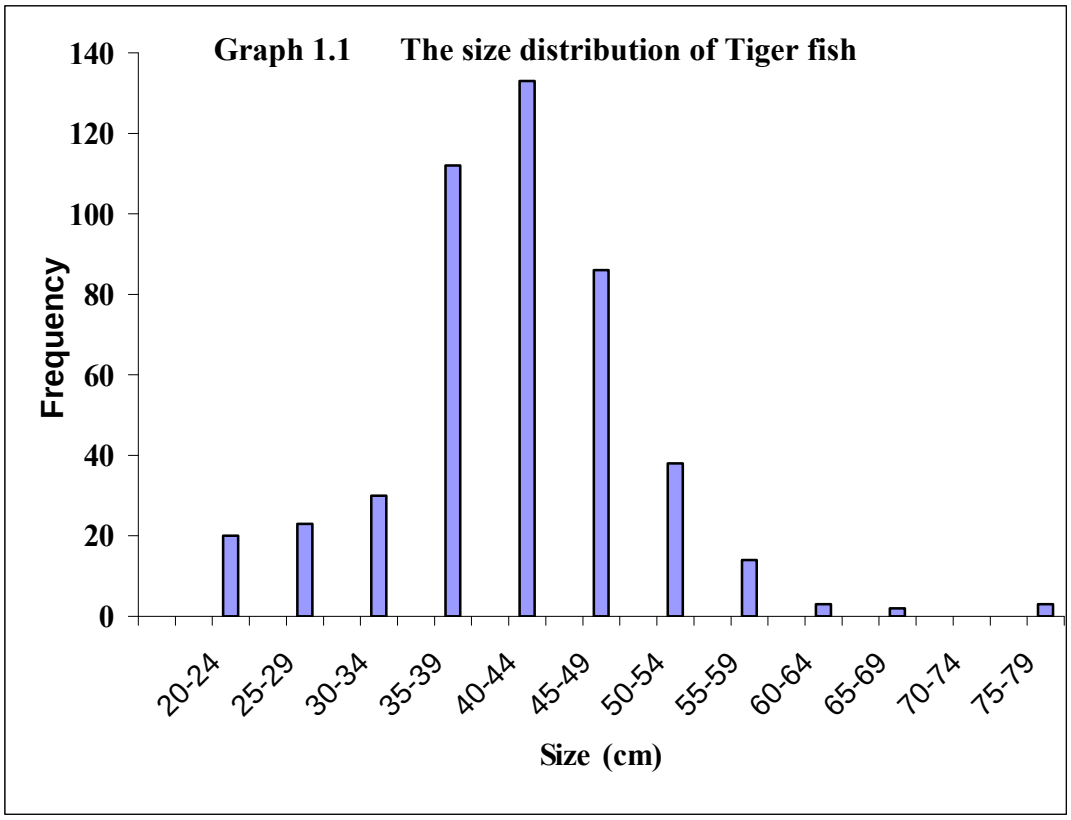
3. A lady grows roses as a hobby. She has six red rose plants and six white rose plants. A friend told her that rose plants produce more flowers when they receive morning sunlight. She reasoned that when rose plants receive morning sunlight instead of afternoon sunlight, they produce more flowers.

   Which plan should she choose to test her friend's idea?

   A. Set all her rose plants in the morning sun. Count the number of roses produced by each plant. Do this for a period of four months. Then find the average number of roses produced by each kind of rose plant.
   B. Set all her rose plants in the morning sunlight for four months. Count the number of flowers produced during this time. Then set all the rose plants in the afternoon sunlight for four months. Count the number of flowers produced during this time.
   C. Set three white rose plants in the morning sunlight and the other three white rose plants in the afternoon sun. Count the number of flowers produced by each white rose plant for four months.
   D. Set three red and three white rose plants in the morning sunlight, and three red and three white rose plants in the afternoon sunlight. Count the number of rose flowers produced by each rose plant for four months.

Questions 4 and 5 refer to the graph below.

The fishery department wants to know the average size of Tiger fish in Tzaneen dam, so that they could prevent over-fishing. They carry out an investigation, and the results of the investigation are presented in the graph below.



4. What is the most common size range of Tiger fish found in Tzaneen dam

A. 75 – 79 cm.
B. 40 – 44 cm.
C. 20 – 79 cm.
D. 45 – 49 cm.

5. In which size range would you find the longest Tiger fish?

A. 75 – 79 cm.
B. 40 – 44 cm.
C. 20 – 79 cm.
D. 35 – 49 cm.

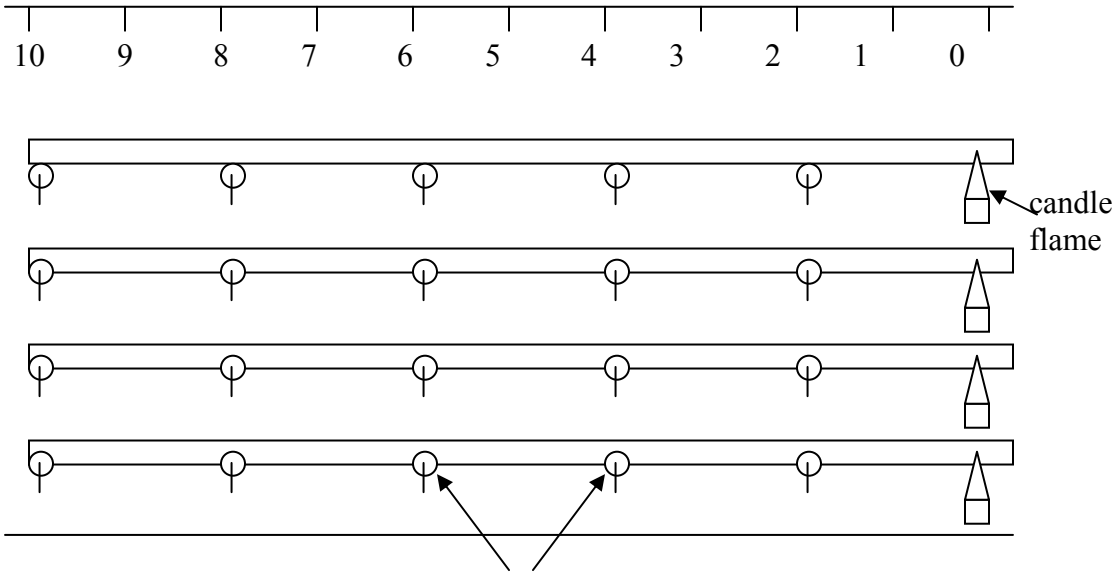
6. Mpho wants to know what determines the time it takes for water to boil. He pours the same amount of water into four containers of different sizes, made of clay, steel, aluminium and copper. He applies the same amount of heat to the containers and measures the time it takes the water in each container to boil.

   Which one of the following could affect the time it takes for water to boil in this investigation?

   A. The shape of the container and the amount water used.
   B. The amount of water in the container and the amount of heat used.
   C. The size and type of the container used.
   D. The type of container and the amount of heat used.

7. A teacher wants to find out how quickly different types of material conduct heat. He uses four rods with the same length and diameter but made of different types of material. He attaches identical pins to the rods using wax, at regular intervals as shown in the diagram below. All the rods were heated on one end at the same time, using candle flames. After two minutes, the pins that fell from each rod were counted.

   Diagram 1.1



How is the speed (rate) of heat conduction by the various rods measured in this study?

A. By determining the rod, which conducted heat faster when heated.
B. By counting the number of pins that fall from each rod after 2 minutes.
C. By counting the number of minutes taken for each pin to fall from the rod.
D. By using wax to measure the rate of heat conduction.

8. A farmer wants to increase the amount of mealies he produces. He decides to study the factors that affect the amount of mealies produced.

Which of the following ideas could he test?

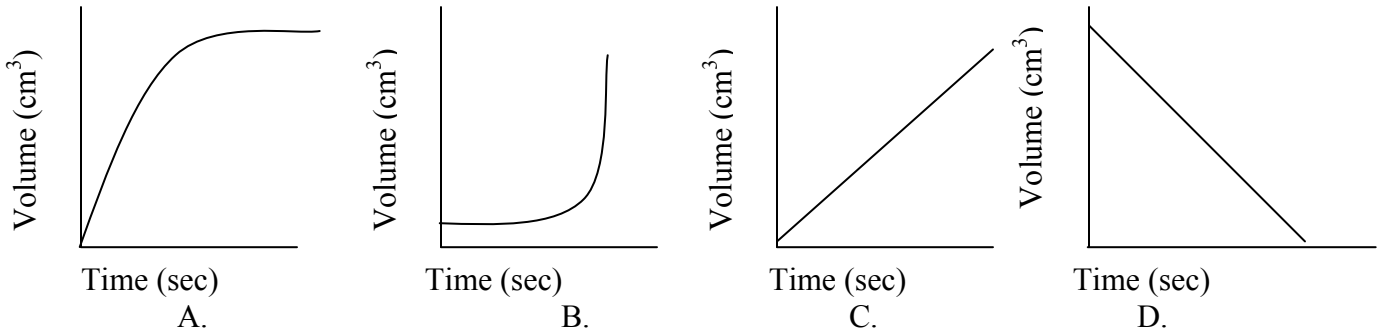
A. The greater the amount of mealies produced, the greater the profit for the year.
B. The greater the amount of fertilizer used, the more the amount of mealies produced.
C. The greater the amount of rainfall, the more effective the fertilizer used will be.
D. The greater the amount of mealies produced, the cheaper the cost of mealies.

9. Sandile carried out an investigation in which she reacted magnesium with dilute hydrochloric acid. She recorded the volume of the hydrogen produced from the reaction, every second. The results are shown below.

Time (seconds)	0	1	2	3	4	5	6	7
Volume (cm^3)	0	14	23	31	38	40	40	40

Table 1.1. Shows the volume of hydrogen produced per second.

Which of the following graphs show these results correctly?



10. A science teacher wanted to find out the effect of exercise on pulse rate. She asked each of three groups of learners to do some push-ups over a given period of time, and then measure their pulse rates: one group did the push-ups for one minute; the second group for two minutes; the third group for three minutes and then a fourth group did not do any push-ups at all.

    How is pulse rate measured in this investigation?

    A. By counting the number of push-ups in one minute.
    B. By counting the number of pulses in one minute.
    C. By counting the number of push-ups done by each group.
    D. By counting the number of pulses per group.

11 Five different hosepipes are used to pump diesel from a tank. The same pump is used for each hosepipe. The following table shows the results of an investigation that was done on the amount of diesel pumped from each hosepipe.

Size (diameter) of hosepipe (mm)	**Amount of diesel pumped per minute (litres)**
8	1
13	2
20	4
26	7
31	12

Table 1.2. Shows the amount of diesel pumped per minute.

Which of the following statements describes the effect of the size of the hosepipe on the amount of diesel pumped per minute?

A. The larger the diameter of the hosepipe, the more the amount of diesel pumped.
B. The more the amount of diesel pumped, the more the time used to pump it.
C. The smaller the diameter of the hosepipe, the higher the speed at which the diesel is pumped.
D. The diameter of the hosepipe has an effect on the amount of diesel pumped.

12. Doctors noticed that if certain bacteria were injected into a mouse, it developed certain symptoms and died. When the cells of the mouse were examined under the microscope, it was seen that the bacteria did not spread through the body of the mouse, but remained at the area of infection. It was therefore thought that the death is not caused by the bacteria but by certain toxic chemicals produced by them.

Which of the statements below provides a possible explanation for the cause of death of the mouse?

A. The mouse was killed by the cells that were removed from it to be examined under the microscope.
B. Bacteria did not spread through the body of the mouse but remained at the site of infection.
C. The toxic chemical produced by the bacteria killed the mouse.
D. The mouse was killed by developing certain symptoms.

13. Thembi thinks that the more the air pressure in a soccer ball, the further it moves when kicked. To investigate this idea, he uses several soccer balls and an air pump with a pressure gauge. How should Thembi test his idea?

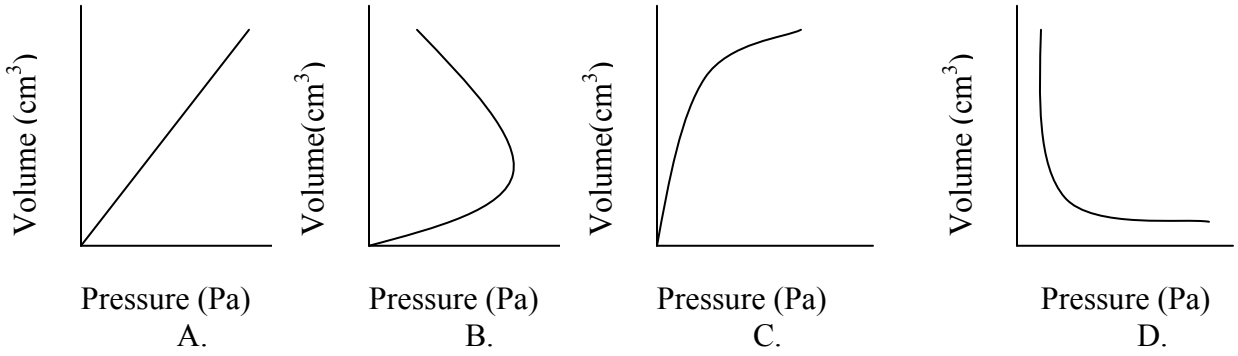
A. Kick the soccer balls with different amounts of force from the same point.
B. Kick the soccer balls having different air pressure from the same point.
C. Kick the soccer balls having the same air pressure at different angles on the ground.
D. Kick the soccer balls having different air pressure from different points on the ground.

14. A science class wanted to investigate the effect of pressure on volume, using balloons. They performed an experiment in which they changed the pressure on a balloon and measured its volume. The results of the experiment are given in the table below.

Pressure on balloon (Pa)	Volume of the balloon (cm^3)
0.35	980
0.70	400
1.03	320
1.40	220
1.72	180

Table 1.3. Shows the relationship between the pressure on a balloon and its volume.

Which of the following graphs represents the above data correctly?



15. A motorist wants to find out if a car uses more fuel when it is driven at high speed. What is the best way of doing this investigation?

A. Ask several drivers how much fuel they use in one hour, when they drive fast, and find the average amount of fuel used per hour.
B. Use his own car to drive several times at different speeds, and he should record the amount of fuel used each time.
C. He must drive his car at high speed, for a week, and then drive it at low speed for another week, and record the amount of fuel used in each case.
D. Ask several drivers to drive different cars covering the same distance many times, at different speeds, and record the amount of fuel used for each trip.

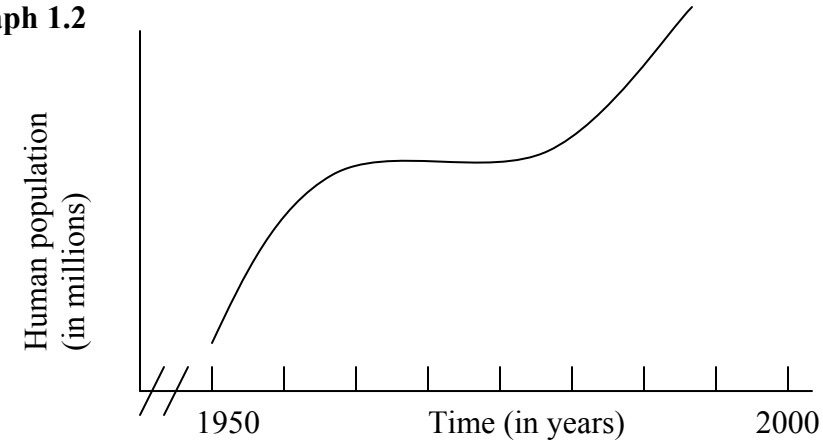
16. A learner observed that anthills (termite moulds) in a certain nature reserve tend to lean towards the west, instead of being straight. In this area, the wind blows towards the direction in which the anthills lean.

Which of the following statements can be tested to determine what causes the anthills to lean towards the west, in this nature reserve?

A. Anthills are made by termites.
B. Anthills lean in the direction in which the wind blows.
C. Anthills lean towards the west to avoid the sun and the rain.
D. The distribution of anthills depends on the direction of the wind.

17. The graph below shows the changes in human population from the year 1950 to 2000.

Graph 1.2



Which of the following statements best describes the graph?

A. The human population increases as the number of years increase.
B. The human population first increases, then it reduces and increases again as the number of years increase.
C. The human population first increases, then it remains the same and increases again as the number of years increase.
D. The human population first increases then it remains the same as the number of years increase.

18. Mulai wants to find out the amount of water contained in meat, cucumber, cabbage and maize grains. She finely chopped each of the foods and carefully measured 10 grams of each. She then put each food in a dish and left all the dishes in an oven set at 100^{o}C. After every 30 minutes interval, she measured the mass of each food, until the mass of the food did not change in two consecutive measurements. She then determined the amount of water contained in each of the foods.

How is the amount of water contained in each food measured in this experiment?

A. By heating the samples at a temperature of 100^{o}C and evaporating the water.
B. By measuring the mass of the foods every 30 minutes and determining the final mass.
C. By finely chopping each food and measuring 10 grams of it, at the beginning of the investigation.
D. By finding the difference between the original and the final mass of each food.

19. In a radio advertisement, it is claimed that Surf produces more foam than other types of powdered soap. Chudwa wanted to confirm this claim. He put the same amount of water in four basins, and added 1 cup of a different type of powdered soap (including surf) to each basin. He vigorously stirred the water in each basin, and observed the one that produced more foam.

Which of the factors below is **NOT** likely to affect the production of foam by powdered soap?

A. The amount of time used to stir the water.
B. The amount of stirring done.
C. The type of basin used.
D. The type of powered soap used.

20 Monde noticed that the steel wool that she uses to clean her pots rusts quickly if exposed to air after using it. She also noticed that it takes a longer time for it to rust if it is left in water. She wondered whether it is the water or the air that causes the wet exposed steel wool to rust.

Which of the following statements could be tested to answer Monde's concern?

A. Steel wool cleans pots better if it is exposed to air.
B. Steel wool takes a longer time to rust if it is left in water.
C. Water is necessary for steel wool to rust.
D. Oxygen can react with steel wool.

21. A science teacher wants to demonstrate the lifting ability of magnets to his learners. He uses many magnets of different sizes and shapes. He weighs the amount of iron filings picked by each magnet.

How is the lifting ability of magnets defined in this investigation?

A. The weight of the iron filings picked up by the magnets.
B. The size of the magnet used.
C. The weight of the magnet used to pick up the iron filings.
D. The shape of the magnet used.

22. Thabo wanted to show his friend that the size of a container affects the rate of water loss, when water is boiled. He poured the same amount of water in containers of different sizes but made of the same material. He applied the same amount of heat to all the containers. After 30 minutes, he measured the amount of water remaining in each container.

How was the rate of water loss measured in this investigation?

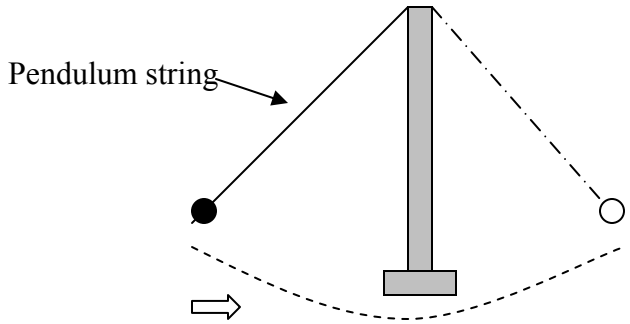
A. By measuring the amount of water in each container after heating it.
B. By using different sizes of the containers to boil the water for 30 minutes.
C. By determining the time taken for the water to boil in each of the containers.
D. By determining the difference between the initial and the final amounts of water, in a given time.

23. A school gardener cuts grass from 7 different football fields. Each week, he cuts a different field. The grass is usually taller in some fields than in others. He makes some guesses about why the height of the grass is different. Which of the following is a suitable testable explanation for the difference in the height of grass.

A. The fields that receive more water have longer grass.
B. Fields that have shorter grass are more suitable for playing football.
C. The more stones there are in the field, the more difficult it is to cut the grass.
D. The fields that absorb more carbon dioxide have longer grass.

24. James wanted to know the relationship between the length of a pendulum string and the time it takes for a pendulum to make a complete swing. He adjusted the pendulum string to different lengths and recorded the time it took the pendulum to make a complete swing.

Diagram 1.2 A pendulum.

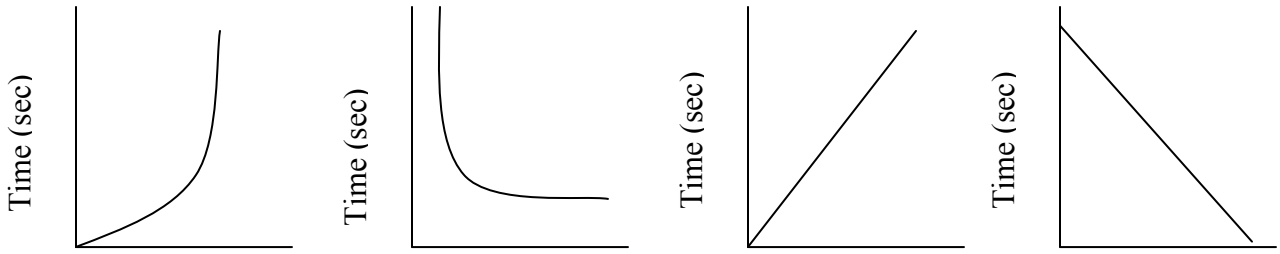


He obtained the following results from an investigation.

Length of string (cm)	80.0	100.0	120.0	140.0	160.0	180.0
Time taken (seconds)	1.80	2.02	2.21	2.39	2.55	2.71

Table 1.4. The relationship between the lengths of a pendulum string and the time the pendulum takes to make a complete swing.

Which of the following graphs represent the above information correctly?



A. B. C. D.

25. A farmer raises chickens in cages. He noticed that some chickens lay more eggs than others. Another farmer tells him that, the amount of food and water given to chicken, and the weight of chicken, affect the number of eggs they lay.

Which of the following is **NOT** likely to be a factor that affects the number of eggs laid by the chickens?

A. The size of the cage where the eggs are laid.
B. The weight of the chickens.
C. The amount of food given to the chickens.
D. The amount of water given to the chickens.

26. A science class wanted to test the factors that might affect plant height. They felt that the following is a list of factors that could be tested: the amount of light, amount of moisture, soil type, and change in temperature.

Which of the statements below could be tested to determine the factor that might affect plant height?

A. An increase in temperature will cause an increase in plant height.
B. An increase in sunlight will cause a decrease in plant moisture.
C. A plant left in light will be greener than one left in the dark.
D. A plant in sand soil loses more water than one in clay soil.

27. A Biology teacher wanted to show her class the relationship between light intensity and the rate of plant growth. She carried out an investigation and got the following results.

Light intensity (Candela)	Plant growth rate (cm)
250	2
800	5
1000	9
1200	11
1800	12
2000	15
2400	13
2800	10
3100	5

Table 1.5. Shows the relationship between light intensity and the growth rate of a plant.

Which of the following statements correctly describes what these results show?

A. As light intensity increases, plant growth also increases.
B. As plant growth increases, light intensity decreases.
C. As plant growth increases, light intensity increases then decreases.
D. As light intensity increases, plant growth increases then decreases.

Questions 28, 29 and 30 refer to the investigation below.

Thabiso is worried about how the cold winter will affect the growth of his tomatoes. He decided to investigate the effect of temperature on the growth rate of tomato plants. He planted tomato seedlings in four identical pots with the same type of soil and the same amount of water. The pots were put in different glass boxes with different temperatures: One at $0^{o}C$, the other at $10^{o}C$, and another at room temperature and the fourth at $50^{o}C$. The growth rates of the tomato plants were recorded at the end of 14 days.

28. What effect does the differences in temperature have in this investigation?

A. The difference in the seasons.
B. The difference in the amount of water used.
C. The difference in growth rates of the tomato plants.
D. The difference in the types of soil used in the different pots.

29. The factor(s) that were being investigated in the above experiment are:

A. Change in temperature and the type of soil used.
B. Change in temperature and the growth rate of the tomato plants.
C. The growth rate of tomato plants and the amount of water used.
D. The type of soil used and the growth rate of the tomato plants.

30. Which of the following factors were kept constant in this investigation?

A. The time and growth rate of tomato plant.
B. The growth rate of tomato plants and the amount of water used.
C. The type of soil and the amount of water used.
D. The temperature and type of soil used.

APPENDIX II.

SCORING KEY FOR THE DEVELOPED TEST INSTRUMENT

Item #	Correct option	Item #	Correct option	Item #	Correct option
1	D	11	A	21	A
2	B	12	C	22	D
3	D	13	B	23	A
4	D	14	D	24	C
5	A	15	D	25	A
6	C	16	C	26	A
7	B	17	C	27	D
8	B	18	D	28	C
9	A	19	C	29	B
10	B	20	C	30	C

APPENDIX III

PERCENTAGE AND NUMBER OF LEARNERS WHO SELECTED EACH OPTION IN THE DIFFERENT PERFORMANCE CATEGORIES.

	Option A						Option B						Option C						Option D					
Qn #	H	%	M	%	L	%	H	%	M	%	L	%	H	%	M	%	L	%	H	%	M	%	L	%
1	8	3.8	37	10	27	13	20	9.6	110	31	70	34	10	4.8	46	13	49	24	**157**	**75**	**149**	**42**	**80**	**38**
2	12	5.8	77	22	54	26	**169**	**81**	**222**	**63**	**83**	**40**	7	3.4	24	6.8	39	19	20	9.6	30	8.5	31	15
3	21	10	25	7.1	53	25	25	12	67	19	48	23	21	10	56	16	37	18	**155**	**75**	**189**	**54**	**65**	**31**
4	9	4.3	57	16	60	29	**162**	**78**	**181**	**51**	**80**	**38**	19	9.1	63	18	52	25	13	6.3	41	12	23	11
5	**131**	**63**	**97**	**27**	**37**	**18**	68	33	207	59	137	66	9	4.3	23	6.5	19	9.1	8	3.8	14	4	13	6.3
6	4	1.9	21	5.9	31	15	24	12	89	25	61	29	**125**	**60**	**113**	**32**	**69**	**33**	44	21	118	33	63	30
7	34	16	50	14	39	19	**105**	**50**	**102**	**29**	**38**	**18**	47	23	87	25	62	30	21	10	97	27	77	37
8	46	22	96	27	39	19	17	8.2	40	11	46	22	8	3.8	25	7.1	31	15	**135**	**65**	**193**	**55**	**86**	**41**
9	7	3.4	11	3.1	30	14	**151**	**73**	**215**	**61**	**87**	**42**	31	15	107	30	71	34	12	5.8	14	4	30	14
10	**150**	**72**	**140**	**40**	**43**	**21**	8	3.8	31	8.8	34	16	39	19	135	38	108	52	9	4.3	31	8.8	31	15
11	13	6.3	70	20	48	23	**108**	**52**	**69**	**20**	**31**	**15**	21	10	85	24	56	27	66	32	125	35	72	35
12	**138**	**66**	**143**	**41**	**48**	**23**	6	2.9	49	14	39	19	3	1.4	52	15	55	26	56	27	111	31	66	32
13	8	3.8	25	7.1	53	25	6	2.9	54	15	50	24	**161**	**77**	**176**	**50**	**94**	**45**	23	11	84	24	38	18
14	65	31	126	36	72	35	**97**	**47**	**101**	**29**	**39**	**19**	26	13	70	20	55	26	11	5.3	58	16	46	22
15	41	20	98	28	71	34	31	15	58	16	61	29	36	17	71	20	48	23	**98**	**47**	**120**	**34**	**28**	**13**
16	11	5.3	51	14	60	29	43	21	76	22	64	31	52	25	141	40	81	39	**93**	**45**	**68**	**19**	**26**	**13**
17	15	7.2	64	18	56	27	**120**	**58**	**108**	**31**	**41**	**20**	22	11	59	17	49	24	50	24	111	31	69	33
18	13	6.3	83	24	71	34	20	9.6	78	22	52	25	**146**	**70**	**144**	**41**	**82**	**39**	9	4.3	25	7.1	39	19
19	31	15	84	24	56	27	20	9.6	78	22	45	22	7	3.4	58	16	69	33	**140**	**67**	**134**	**38**	**41**	**20**
20	13	6.3	68	19	37	18	8	3.8	45	13	37	18	**159**	**76**	**113**	**32**	**70**	**34**	10	4.8	122	35	81	39
21	45	22	106	30	67	32	8	3.8	77	22	50	24	**114**	**55**	**93**	**26**	**51**	**25**	26	13	78	22	47	23
22	**123**	**59**	**108**	**31**	**39**	**19**	41	20	140	40	61	29	16	7.7	35	9.9	27	13	22	11	90	25	62	30
23	47	23	83	24	60	29	44	21	103	29	82	39	20	9.6	69	20	39	19	**95**	**46**	**90**	**25**	**29**	**14**
24	**126**	**61**	**175**	**50**	**70**	**34**	41	20	87	25	47	23	19	9.1	45	13	51	25	21	10	29	8.2	50	24
25	38	18	60	17	53	25	26	13	76	22	49	24	**127**	**61**	**144**	**41**	**86**	**41**	15	7.2	50	14	37	18
26	**176**	**85**	**182**	**52**	**54**	**26**	20	9.6	75	21	65	31	10	4.8	61	17	57	27	4	1.9	22	6.2	20	9.6
27	**92**	**44**	**63**	**18**	**32**	**15**	37	18	99	28	77	37	42	20	73	21	61	29	35	17	95	27	52	25
28	24	12	81	23	58	28	18	8.7	58	16	55	26	27	13	54	15	48	23	**137**	**66**	**145**	**41**	**48**	**23**
29	65	31	119	34	90	43	11	5.3	45	13	36	17	**99**	**48**	**104**	**29**	**69**	**33**	15	7.2	60	17	43	21
30	9	4.3	50	14	47	23	**141**	**68**	**123**	**35**	**36**	**17**	43	21	110	31	88	42	11	5.3	48	14	47	23
31	28	13	88	25	49	24	8	3.8	64	18	40	19	**132**	**63**	**93**	**26**	**72**	**35**	30	14	91	26	67	32

APPENDIX IV

COMPLETE ITEM RESPONSE PATTERN FROM THE MAIN STUDY

Option	A				B				C				D				E				
Qn #	H	M	L	Tot	H	M	L	Tot	H	M	L	Tot	H	M	L	Tot	H	M	L	Tot	G. Tot
1	8	37	27	72	20	110	70	200	10	46	49	105	**157**	**149**	**80**	386	1	4	1	6	769
2	12	77	54	143	**169**	**222**	**83**	474	7	24	39	70	20	30	31	81	0	1	0	1	769
3	21	25	53	99	25	67	48	140	21	56	37	114	**155**	**189**	**65**	409	0	5	2	7	769
4	9	57	60	126	**162**	**181**	**80**	423	19	63	52	134	13	41	23	77	2	3	4	9	769
5	**131**	**97**	**37**	265	68	207	137	412	9	23	19	51	8	14	13	35	0	3	3	6	769
6	4	21	31	56	24	89	61	174	**125**	**113**	**69**	307	44	118	63	225	2	3	2	7	769
7	34	50	39	123	**105**	**102**	**38**	245	47	87	62	196	21	97	77	195	1	7	2	10	769
8	46	96	39	181	17	40	46	103	8	25	31	64	**135**	**193**	**86**	414	1	2	4	7	769
9	7	11	30	48	**151**	**215**	**87**	453	31	107	71	209	12	14	30	56	0	2	1	3	769
10	**150**	**140**	**43**	333	8	31	34	73	39	135	108	282	9	31	31	71	2	4	4	10	769
11	13	70	48	131	**108**	**69**	**31**	208	21	85	56	162	66	125	72	263	0	3	2	5	769
12	**138**	**143**	**48**	329	6	49	39	94	3	52	55	110	56	111	66	233	0	1	2	3	769
13	8	25	53	82	6	54	50	110	**161**	**176**	**94**	431	23	84	38	145	0	1	0	1	769
14	65	126	72	263	**97**	**101**	**39**	237	26	70	55	151	11	58	46	115	0	3	0	3	769
15	41	98	71	210	31	58	61	150	36	71	48	155	**98**	**120**	**28**	246	2	1	5	8	769
16	11	51	60	122	43	76	64	183	52	141	81	274	**93**	**68**	**26**	187	0	1	2	3	769
17	15	64	56	135	**120**	**108**	**41**	269	22	59	49	130	50	111	69	230	1	2	2	5	769
18	13	83	71	167	20	78	52	150	**146**	**144**	**82**	372	9	25	39	73	1	3	3	7	769
19	31	84	56	171	20	78	45	143	7	58	69	134	**140**	**134**	**41**	315	0	2	4	6	769
20	13	68	37	118	8	45	37	90	**159**	**113**	**70**	342	10	122	81	213	2	1	3	6	769
21	45	106	67	218	8	77	50	135	**114**	**93**	**51**	258	26	78	47	151	0	5	2	7	769
22	**123**	**108**	**39**	270	41	140	61	242	16	35	27	78	22	90	62	174	0	3	2	5	769
23	47	83	60	190	44	103	82	229	20	69	39	128	**95**	**90**	**29**	214	2	4	2	8	769
24	**126**	**175**	**70**	371	41	87	47	175	19	45	51	115	21	29	50	100	1	6	1	8	769
25	38	60	53	151	26	76	49	151	**127**	**144**	**86**	357	15	50	37	102	2	3	3	8	769
26	**176**	**182**	**54**	412	20	75	65	160	10	61	57	128	4	22	20	46	3	9	11	23	769
27	**92**	**63**	**32**	187	37	99	77	213	42	73	61	176	35	95	52	182	2	4	5	11	769
28	24	81	58	163	18	58	55	131	27	54	48	129	**137**	**145**	**48**	330	2	10	4	16	769
29	65	119	90	274	11	45	36	92	**99**	**104**	**69**	272	15	60	43	118	4	7	2	13	769
30	9	50	47	106	**141**	**123**	**36**	300	43	110	88	241	11	48	47	106	4	8	4	16	769
31	28	88	49	165	8	64	40	112	**132**	**93**	**72**	297	30	91	67	188	2	2	3	7	769

KEY: Qn # = item number

H = number of high scorers who selected the option

M = number of medium scorers who selected the option

L = number of low scorers who selected the option

Tot = the total number of learners who selected the option

G. Tot = the total number of learners whop wrote the test.

Option E =number of learners who omitted or selected more than one option to a question

APPENDIX V
ITEM RESPONSE PATTERN ACCORDING TO THE SCIENCE PROCESS SKILLS MEASURED (IN PERCENTAGE).

	A			B			C			D			E			
Item number	H	M	L	H	M	L	H	M	L	H	M	L	H	M	L	Total
A Identifying and controlling variables																
2	5.8	22	26	81	63	40	3.4	6.8	19	9.6	8.3	15	0	0.3	0	100
6	1.9	5.9	15	12	25	29	60	32	33	21	33	30	1	0.8	1	100
20	6.3	19	18	3.8	13	18	76	32	34	4.8	35	39	1	0.3	1.4	101
26	85	52	26	9.6	21	31	4.8	17	27	1.9	6.2	9.6	1.4	2.5	5.3	100
29	31	34	43	5.3	13	17	48	29	33	7.2	17	21	1.9	2	1	101
30	4.3	14	23	68	35	17	21	31	42	5.3	14	23	1.9	2.3	1.9	101
31	13	25	24	3.8	18	19	63	26	35	14	26	32	1	0.6	1.4	101
B Stating hypotheses																
9	3.4	3.1	14	73	61	42	15	30	34	5.8	4	14	0	0.6	0.5	100
13	3.8	7.1	25	2.9	15	24	77	50	45	11	24	18	0	0.3	0	101
17	7.2	18	27	58	31	20	11	17	24	24	31	33	0.5	0.6	1	101
21	22	30	32	3.8	22	28	55	26	25	4.3	22	23	0	1.4	1	99
24	61	50	34	20	23	23	9.1	13	25	10	8.2	24	0.5	1.7	0.5	101
27	44	18	15	18	28	37	20	21	29	17	27	25	1	1.1	2.4	101
C Operational definitions																
1	3.8	10	13	9.6	31	34	4.8	13	24	75	42	38	0.5	1.1	0.5	100
7	16	14	19	50	29	18	23	25	30	10	27	37	0.5	2	1	101
11	6.3	20	23	52	20	15	10	24	27	32	35	35	0	0.8	1	100
19	15	24	27	9.6	22	22	3.4	16	33	67	38	20	0	0.6	0.9	100
22	59	31	19	20	40	29	7.7	9.9	13	11	25	30	0	0.8	1	99
23	23	24	29	21	29	39	9.6	20	19	46	25	14	1	1.1	1	101
D Graphing and interpreting data																
4	4.3	16	29	78	51	39	9.1	18	25	6.3	12	11	1	0.8	1.9	101
5	63	27	18	33	59	66	4.3	6.5	9.1	3.8	4	6.3	0	0.8	1.4	101
8	22	27	19	8.2	11	22	3.8	7.1	15	65	55	41	0.5	0.6	1.9	100
10	72	40	21	3.8	8.8	16	19	38	52	4.3	8.8	15	1	1.1	1.9	101
12	66	41	23	2.9	14	19	1.4	15	26	27	31	32	0	0.3	1	100
15	20	28	34	15	16	29	17	20	23	47	34	13	1	0.3	2.4	100
18	6.3	24	34	9.6	22	25	70	41	39	4.3	7.1	19	0.5	0.8	1.4	101
25	18	17	25	13	22	24	61	41	41	7.1	14	18	1	0.8	1.4	101
28	12	23	28	8.7	16	26	13	15	23	66	41	23	1	2.8	1.9	100
E Experimental design																
3	10	7.1	25	12	19	23	10	16	18	75	54	31	0	1.4	1	101
14	31	36	35	47	29	19	13	20	26	5.3	16	22	0	0.8	0	100
16	5.3	14	29	21	22	31	25	36	39	45	19	13	0	0.3	1	100

APPENDIX VI

ITEM RESPONSE PATTERN ACCORDING TO GRADELEVELS

KEY
Options A =1; B = 2 C = 3; D = 4; E = ERROR.
I = Item number
R = Correct response for the item.
H = High scorers.
M = Medium scorers.
L = Low scorers.
n = total number of learners from the category.
N = total number of learners who wrote the test in the grade.

(a) GRADE 9

		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31
H	*R*	***4***	***2***	***4***	***2***	***1***	***3***	***2***	***4***	***2***	***1***	***2***	***1***	***3***	***2***	***4***	***4***	***2***	***3***	***4***	***3***	***3***	***1***	***4***	***1***	***3***	***1***	***1***	***4***	***3***	***2***	***3***
1	**A**	2	4	31	4	**43**	3	16	23	1	**44**	7	**38**	5	24	20	7	7	14	14	8	21	**30**	19	**42**	16	**57**	**27**	16	29	5	12
2	**B**	7	**59**	11	**57**	22	12	**31**	7	**53**	3	**36**	4	1	**32**	10	13	**35**	12	14	8	3	26	21	13	9	8	15	7	7	**38**	4
3	**C**	5	2	8	6	1	**36**	14	3	13	22	8	2	**50**	10	20	21	9	**40**	5	**40**	**39**	5	7	6	**38**	3	17	10	**20**	18	**36**
4	**D**	**57**	6	**51**	3	5	19	9	**38**	4	1	20	27	15	5	**21**	**30**	20	4	**38**	14	8	10	**22**	10	7	2	12	**37**	12	8	17
5	**E**	0	0	0	1	0	1	1	0	0	1	0	0	0	0	0	0	0	1	0	1	0	0	2	0	1	1	0	1	3	2	2
	n	**71**	**71**	**71**	**71**	**71**	**71**	**71**	**71**	**71**	**71**	**71**	**71**	**71**	**71**	**71**	**71**	**71**	**71**	**71**	**71**	**71**	**71**	**71**	**71**	**71**	**71**	**71**	**71**	**71**	**71**	**71**
M	*R*	***4***	***2***	***4***	***2***	***1***	***3***	***2***	***4***	***2***	***1***	***2***	***1***	***3***	***2***	***4***	***4***	***2***	***3***	***4***	***3***	***3***	***1***	***4***	***1***	***3***	***1***	***1***	***4***	***3***	***2***	***3***
	A	9	28	19	26	**46**	12	23	33	8	**38**	30	**43**	18	43	45	31	29	46	25	15	28	**14**	36	**62**	25	**59**	**30**	36	49	26	34
	B	42	**71**	29	**60**	56	39	**35**	18	**68**	15	**20**	23	19	**30**	23	26	**34**	20	32	23	26	48	36	33	24	26	39	27	18	**36**	29
	C	7	10	31	17	13	**37**	22	11	41	56	37	22	**48**	25	26	45	25	**42**	23	**26**	**39**	20	27	18	**50**	32	24	19	**31**	39	**17**
	D	**63**	13	**42**	18	6	33	39	**60**	5	11	34	34	36	22	**28**	**20**	34	12	**42**	58	27	40	**22**	8	18	4	27	**39**	23	18	39
	E	1	0	1	1	1	1	3	0	0	2	0	0	1	2	0	0	0	2	0	0	2	0	1	1	5	1	2	1	1	3	3
	n	**122**	**122**	**122**	**122**	**122**	**122**	**122**	**122**	**122**	**122**	**121**	**122**	**122**	**122**	**122**	**122**	**122**	**122**	**122**	**122**	**122**	**122**	**122**	**122**	**122**	**122**	**122**	**122**	**122**	**122**	**122**
L	*R*	***4***	***2***	***4***	***2***	***1***	***3***	***2***	***4***	***2***	***1***	***2***	***1***	***3***	***2***	***4***	***4***	***2***	***3***	***4***	***3***	***3***	***1***	***4***	***1***	***3***	***1***	***1***	***4***	***3***	***2***	***3***
	A	11	19	20	26	**7**	17	19	11	8	**9**	21	**17**	22	27	30	27	19	29	14	10	18	**8**	24	**20**	19	**18**	**10**	24	29	20	14
	B	17	**31**	15	**29**	49	21	**16**	17	**24**	14	**8**	14	18	**9**	21	23	**15**	19	19	14	23	22	28	14	16	22	21	22	17	**5**	20
	C	10	13	16	9	11	**13**	9	14	26	34	18	23	**14**	15	13	13	15	**11**	27	**9**	**12**	14	17	20	**23**	23	15	16	**9**	26	**11**
	D	**32**	8	**19**	5	4	20	26	**27**	12	12	24	17	17	20	**5**	**8**	21	12	**10**	37	16	26	**2**	17	13	8	22	**8**	14	19	26
	E	1	0	1	2	0	0	1	2	1	2	0	0	0	0	2	0	1	0	1	1	2	1	0	0	0	0	3	1	2	1	0
	n	**71**	**71**	**71**	**71**	**71**	**71**	**71**	**71**	**71**	**71**	**71**	**71**	**71**	**71**	**71**	**71**	**71**	**71**	**71**	**71**	**71**	**71**	**71**	**71**	**71**	**71**	**71**	**71**	**71**	**71**	**71**
N		**264**	**264**	**264**	**264**	**264**	**264**	**264**	**264**	**264**	**264**	**264**	**264**	**264**	**264**	**264**	**264**	**264**	**264**	**264**	**264**	**264**	**264**	**264**	**264**	**264**	**264**	**264**	**264**	**264**	**264**	**264**

(b) GRADE 10

	I	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31
H	*R*	*4*	*2*	*4*	*2*	*1*	*3*	*2*	*4*	*2*	*1*	*2*	*1*	*3*	*2*	*4*	*4*	*2*	*3*	*4*	*3*	*3*	*1*	*4*	*1*	*3*	*1*	*1*	4	3	2	3
	A	4	4	4	2	**43**	0	11	10	4	**43**	3	**49**	1	22	13	2	3	6	5	1	26	**56**	24	**33**	12	**59**	**34**	5	25	3	15
	B	9	**58**	6	**54**	24	6	**34**	8	**47**	5	**36**	1	2	**34**	14	15	**38**	9	11	0	0	2	10	18	12	5	12	8	3	**45**	3
	C	5	3	8	9	0	**41**	15	3	16	12	8	1	**57**	10	10	22	9	**52**	0	**65**	**31**	4	5	8	**39**	1	12	8	**38**	18	**32**
	D	**50**	4	**51**	4	2	21	9	**48**	2	8	22	18	9	3	**31**	**30**	18	2	**53**	3	12	7	**30**	9	5	2	9	**47**	2	1	17
	E	1	0	0	0	0	1	0	0	0	1	0	0	0	0	1	0	1	0	0	0	0	0	0	1	1	2	2	1	1	2	2
	n	**69**	**69**	**69**	**69**	**69**	**69**	**69**	**69**	**69**	**69**	**69**	**69**	**69**	**69**	**69**	**69**	**69**	**69**	**69**	**69**	**69**	**69**	**69**	**69**	**69**	**69**	**69**	**69**	**69**	**69**	**69**
M	***R***	***4***	***2***	***4***	***2***	***1***	***3***	***2***	***4***	***2***	***1***	***2***	***1***	***3***	***2***	***4***	***4***	***2***	***3***	***4***	***3***	***3***	***1***	***4***	***1***	***3***	***1***	***1***	*4*	*3*	*2*	*3*
	A	12	29	8	18	**37**	7	13	36	4	**39**	22	**46**	11	48	40	15	20	36	28	25	34	**44**	27	**51**	22	**55**	**20**	31	46	17	30
	B	34	**77**	25	**67**	67	30	**38**	8	**70**	11	**26**	13	12	**30**	22	21	**42**	30	21	15	29	43	38	30	23	24	21	19	14	**36**	21
	C	12	3	13	18	5	**37**	32	9	38	49	26	16	**69**	23	25	51	18	**39**	16	**38**	**25**	10	22	17	**48**	19	30	15	**28**	37	**25**
	D	**56**	7	**69**	12	6	42	31	**63**	4	18	40	42	25	16	**30**	**30**	37	11	**51**	38	27	18	**28**	15	18	11	36	**44**	21	18	32
	E	3	1	2	2	2	1	3	1	1	0	3	0	0	0	0	0	0	1	1	1	2	2	2	4	6	8	10	8	8	9	9
	n	**117**	**117**	**117**	**117**	**117**	**117**	**117**	**117**	**117**	**117**	**117**	**117**	**117**	**117**	**117**	**117**	**117**	**117**	**117**	**117**	**117**	**117**	**117**	**117**	**117**	**117**	**117**	**117**	**117**	**117**	**117**
L	***R***	***4***	***2***	***4***	***2***	***1***	***3***	***2***	***4***	***2***	***1***	***2***	***1***	***3***	***2***	***4***	***4***	***2***	***3***	***4***	***3***	***3***	***1***	***4***	***1***	***3***	***1***	***1***	*4*	*3*	*2*	*3*
	A	11	17	18	19	**19**	7	10	13	10	**12**	18	**18**	19	25	23	19	23	27	19	15	28	**22**	18	**19**	20	**21**	**13**	17	25	14	16
	B	22	**31**	20	**26**	42	18	**10**	15	**30**	10	**10**	10	15	**13**	19	17	**13**	17	12	19	15	17	32	11	17	19	23	16	12	**11**	15
	C	12	13	9	15	2	**17**	20	12	17	31	16	18	**21**	18	15	23	18	**8**	23	**12**	**10**	8	7	21	**17**	24	18	12	**15**	28	**15**
	D	**24**	8	**21**	9	4	26	29	**27**	12	14	23	22	14	13	**10**	**8**	14	15	**13**	22	16	22	**10**	17	14	4	14	23	17	14	23
	E	0	0	1	0	2	1	0	2	0	2	2	1	0	0	2	2	1	2	2	1	0	0	2	0	1	1	1	1	0	2	0
	n	**69**	**69**	**69**	**69**	**69**	**69**	**69**	**69**	**69**	**69**	**69**	**69**	**69**	**69**	**69**	**69**	**69**	**69**	**69**	**69**	**69**	**69**	**69**	**68**	**69**	**69**	**69**	**69**	**69**	**69**	**69**
	N	**255**	**255**	**255**	**255**	**255**	**255**	**255**	**255**	**255**	**255**	**255**	**255**	**255**	**255**	**255**	**255**	**255**	**255**	**255**	**255**	**255**	**255**	**255**	**255**	**255**	**255**	**255**	**255**	**255**	**255**	**255**

(c) GRADE 11

	I	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31
H	R	*4*	*2*	*4*	*2*	*1*	*3*	*2*	*4*	*2*	*1*	*2*	*1*	*3*	*2*	*4*	*4*	*2*	*3*	*4*	*3*	*3*	*1*	*4*	*1*	*3*	*1*	*1*	*4*	*3*	*2*	*3*
	A	2	4	1	3	**45**	1	7	14	2	**63**	3	**51**	2	28	8	2	5	3	12	5	13	**37**	4	**51**	10	**60**	**31**	3	25	1	3
	B	13	**52**	9	**51**	22	6	**40**	2	**51**	0	**36**	1	3	**31**	7	15	**47**	8	5	5	5	19	13	10	5	7	10	3	1	**58**	1
	C	3	2	5	7	0	**48**	18	2	9	5	5	0	**54**	6	6	18	4	**54**	2	**54**	**44**	7	8	5	**50**	1	13	9	**41**	7	**59**
	D	**50**	10	**53**	6	1	13	3	**49**	6	0	24	16	9	3	**46**	**33**	12	3	**49**	3	6	5	**43**	2	3	0	14	**53**	1	2	5
	E	0	0	0	1	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
	n	**68**	**68**	**68**	**68**	**68**	**68**	**68**	**68**	**68**	**68**	**68**	**68**	**68**	**68**	**68**	**68**	**68**	**68**	**68**	**68**	**68**	**68**	**68**	**68**	**68**	**68**	**68**	**68**	**68**	**68**	**68**
M	*R*	*4*	*2*	*4*	*2*	*1*	*3*	*2*	*4*	*2*	*1*	*2*	*1*	*3*	*2*	*4*	*4*	*2*	*3*	*4*	*3*	*3*	*1*	*4*	*1*	*3*	*1*	*1*	*4*	*3*	*2*	*3*
	A	13	19	9	21	**23**	7	24	24	3	**62**	26	**51**	9	30	18	21	24	21	28	18	38	**20**	29	**73**	25	**72**	**23**	19	39	16	28
	B	44	**74**	13	**54**	84	20	**29**	14	**77**	5	**23**	13	23	**41**	13	29	**32**	28	25	11	22	56	29	24	29	25	39	12	13	**51**	14
	C	27	11	12	28	5	**39**	33	5	28	43	22	14	**59**	22	20	45	16	**63**	19	**49**	**29**	5	15	10	**46**	10	19	20	**45**	34	**51**
	D	**30**	10	**78**	11	2	47	27	**70**	5	2	43	35	23	20	**62**	**18**	40	2	**41**	36	24	32	**40**	6	14	7	32	**62**	16	12	20
	E	0	0	2	0	0	1	1	1	1	2	0	1	0	1	1	1	2	0	1	0	1	1	1	1	0	0	1	1	1	1	1
	n	**114**	**114**	**114**	**114**	**114**	**114**	**114**	**114**	**114**	**114**	**114**	**114**	**114**	**114**	**114**	**114**	**114**	**114**	**114**	**114**	**114**	**114**	**114**	**114**	**114**	**114**	**114**	**114**	**114**	**114**	**114**
L	*R*	*4*	*2*	*4*	*2*	*1*	*3*	*2*	*4*	*2*	*1*	*2*	*1*	*3*	*2*	*4*	*4*	*2*	*3*	*4*	*3*	*3*	*1*	*4*	*1*	*3*	*1*	*1*	*4*	*3*	*2*	*3*
	A	5	18	17	17	**11**	7	10	15	12	**22**	9	**13**	12	20	16	14	14	25	23	12	21	**9**	14	**31**	14	**15**	**9**	17	37	13	16
	B	31	**21**	13	**25**	46	22	**12**	14	**33**	10	**13**	15	17	**17**	21	24	**13**	16	17	9	18	32	22	22	26	24	23	17	7	**20**	14
	C	8	14	13	15	5	**21**	24	7	17	31	27	12	**22**	18	17	20	21	**14**	9	**14**	**13**	5	15	8	**14**	21	19	16	**12**	20	**18**
	D	**24**	15	**25**	9	5	17	21	**32**	6	5	19	27	17	13	**13**	**10**	20	12	**18**	32	16	21	**17**	6	12	8	16	**17**	12	14	20
	E	0	0	0	2	1	1	1	0	0	0	0	1	0	0	1	0	0	1	1	1	0	1	0	1	2	0	1	1	0	1	0
	n	**68**	**68**	**68**	**68**	**68**	**68**	**68**	**68**	**68**	**68**	**68**	**68**	**68**	**68**	**68**	**68**	**68**	**68**	**68**	**68**	**68**	**68**	**68**	**68**	**68**	**68**	**68**	**68**	**68**	**68**	**68**
	N	**250**	**250**	**250**	**250**	**250**	**250**	**250**	**250**	**250**	**250**	**250**	**250**	**250**	**250**	**250**	**250**	**250**	**250**	**250**	**250**	**250**	**250**	**250**	**250**	**250**	**250**	**250**	**250**	**250**	**250**	**250**

Scoring Key

I	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31
R	*4*	*2*	*4*	*2*	*1*	*3*	*2*	*4*	*2*	*1*	*2*	*1*	*3*	*2*	*4*	*4*	*2*	*3*	*4*	*3*	*3*	*1*	*4*	*1*	*3*	*1*	*1*	*4*	*3*	*2*	*3*

APPENDIX VII

DISCRIMINATION AND DIFFICULTY INDICES FOR EACH ITEM, ACCORDING TO GRADE LEVELS.

KEY:

Item No;	The number of the item in the test instrument.
High;	The number of high scorers who selected the correct option.
Med.	The number of medium scorers who selected the correct option.
Low;	The number of low scorers who selected the correct option.
n;	The total number of learners who selected the correct response.
N;	The total number of learners who wrote the test and were considered for the analysis.
%nH;	The percentage of high scorers who selected the correct option.
%nM;	The percentage of medium scorers who selected the correct option.
%nL;	The percentage of low scorers who selected the correct option.
%n;	The percentage of the total number of learners who selected the correct option.
Discrimin;	The discrimination index for the item.
Difficulty	The index of difficulty for the item.

(A) GRADE 9

Item No.	High	Med.	Low	n	% nH	% nM	% nL	% n	Discrimin	Difficulty
1	59	63	32	154	83.09859	48.36066	83.09859	58.33333	0.380282	0.583333
2	59	71	31	161	83.09859	48.36066	83.09859	60.98485	0.394366	0.609848
3	51	42	19	112	71.83099	41.80328	71.83099	42.42424	0.450704	0.424242
4	57	60	29	146	71.83099	41.80328	71.83099	55.30303	0.394366	0.553030
5	43	46	7	96	60.56338	35.2459	60.56338	36.36364	0.507042	0.363636
6	36	37	13	86	50.70423	29.5082	50.70423	32.57576	0.323944	0.325758
7	31	35	16	82	43.66197	25.40984	43.66197	31.06061	0.211268	0.310606
8	38	60	27	125	53.52113	31.14754	53.52113	47.34848	0.15493	0.473485
9	53	68	24	145	74.64789	43.44262	74.64789	54.92424	0.408451	0.549242
10	44	38	9	91	61.97183	36.06557	61.97183	34.4697	0.492958	0.344697
11	36	20	8	64	50.70423	29.5082	50.70423	24.24242	0.394366	0.242424
12	38	43	17	98	53.52113	31.14754	53.52113	37.12121	0.295775	0.371212
13	50	48	14	112	70.42254	40.98361	70.42254	42.42424	0.507042	0.424242
14	32	30	9	71	45.07042	26.22951	45.07042	26.89394	0.323944	0.268939
15	21	28	5	54	29.57746	17.21311	29.57746	20.45455	0.225352	0.204545
16	30	20	8	58	42.25352	24.59016	42.25352	21.9697	0.309859	0.219697
17	35	34	15	84	49.29577	28.68852	49.29577	31.81818	0.28169	0.318182
18	40	42	11	93	56.33803	32.78689	56.33803	35.22727	0.408451	0.352273
19	38	42	10	90	53.52113	31.14754	53.52113	34.09091	0.394366	0.340909
20	40	26	9	75	56.33803	32.78689	56.33803	28.40909	0.43662	0.284091
21	39	39	12	90	54.92958	31.96721	54.92958	34.09091	0.380282	0.340909
22	30	14	8	52	42.25352	24.59016	42.25352	19.69697	0.309859	0.19697
23	22	22	2	46	30.98592	18.03279	30.98592	17.42424	0.28169	0.174242
24	42	62	20	124	59.15493	34.42623	59.15493	46.9697	0.309859	0.469697
25	38	50	23	111	53.52113	31.14754	53.52113	42.04545	0.211268	0.420455
26	57	59	18	134	80.28169	46.72131	80.28169	50.75758	0.549296	0.507576
27	27	30	10	67	38.02817	22.13115	38.02817	25.37879	0.239437	0.253788

28	37	39	8	84	52.11268	30.32787	52.11268	31.81818	0.408451	0.318182
29	20	31	9	60	28.16901	16.39344	28.16901	22.72727	0.15493	0.227273
30	38	36	5	79	53.52113	31.14754	53.52113	29.92424	0.464789	0.299242
31	36	17	11	64	50.70423	29.5082	50.70423	24.24242	0.352113	0.242424
Mean	**39.26**	**40.39**	**14.16**	**93.81**	**55.020**	**32.020**	**55.020**	**35.533**	**0.353476**	**0.355327**

(B) GRADE 10

Item No.	**High**	**Med.**	**Low**	**n**	**% nH**	**% nM**	**% nL**	**% n**	**Discrimin**	**Difficulty**
1	50	56	24	130	72.46377	45.90164	34.78261	50.98039	0.376812	0.509804
2	58	77	31	166	84.05797	63.11475	44.92754	65.09804	0.391304	0.65098
3	51	69	21	141	73.91304	56.55738	30.43478	55.29412	0.434783	0.552941
4	54	67	26	147	78.26087	54.91803	37.68116	57.64706	0.405797	0.576471
5	43	37	19	99	62.31884	30.32787	27.53623	38.82353	0.347826	0.388235
6	41	37	17	95	59.42029	30.32787	24.63768	37.2549	0.347826	0.372549
7	34	38	10	82	49.27536	31.14754	14.49275	32.15686	0.347826	0.321569
8	48	63	27	138	69.56522	51.63934	39.13043	54.11765	0.304348	0.541176
9	47	70	30	147	68.11594	57.37705	43.47826	57.64706	0.246377	0.576471
10	43	39	12	94	62.31884	31.96721	17.3913	36.86275	0.449275	0.368627
11	36	26	10	72	52.17391	21.31148	14.49275	28.23529	0.376812	0.282353
12	49	46	18	113	71.01449	37.70492	26.08696	44.31373	0.449275	0.443137
13	57	69	21	147	82.6087	56.55738	30.43478	57.64706	0.521739	0.576471
14	34	30	13	77	49.27536	24.59016	18.84058	30.19608	0.304348	0.301961
15	31	30	10	71	44.92754	24.59016	14.49275	27.84314	0.304348	0.278431
16	30	30	8	68	43.47826	24.59016	11.5942	26.66667	0.318841	0.266667
17	38	42	13	93	55.07246	34.42623	18.84058	36.47059	0.362319	0.364706
18	52	39	8	99	75.36232	31.96721	11.5942	38.82353	0.637681	0.388235
19	53	51	13	117	76.81159	41.80328	18.84058	45.88235	0.57971	0.458824
20	65	38	12	115	94.2029	31.14754	17.3913	45.09804	0.768116	0.45098
21	31	25	10	66	44.92754	20.4918	14.49275	25.88235	0.304348	0.258824
22	56	44	22	122	81.15942	36.06557	31.88406	47.84314	0.492754	0.478431
23	30	28	10	68	43.47826	22.95082	14.49275	26.66667	0.289855	0.266667
24	33	51	19	103	47.82609	41.80328	27.53623	40.39216	0.202899	0.403922
25	39	48	17	104	56.52174	39.34426	24.63768	40.78431	0.318841	0.407843
26	59	55	21	135	85.50725	45.08197	30.43478	52.94118	0.550725	0.529412
27	34	20	13	67	49.27536	16.39344	18.84058	26.27451	0.304348	0.262745
28	47	44	23	114	68.11594	36.06557	33.33333	44.70588	0.347826	0.447059
29	38	28	15	81	55.07246	22.95082	21.73913	31.76471	0.333333	0.317647
30	45	36	11	92	65.21739	29.5082	15.94203	36.07843	0.492754	0.360784
31	37	25	10	72	53.62319	20.4918	14.49275	28.23529	0.391304	0.282353
Mean	**43.967**	**43.806**	**16.581**		**63.7213**	**35.9069**	**24.0299**	**40.9237**	**0.39691**	**0.40923**

(C) GRADE 11

Item No.	High	Med.	Low	n	% nH	% nM	% nL	% n	Discrimin	Difficulty
1	50	30	24	104	73.52941	25.21008	35.29412	40	0.382353	0.416
2	52	74	21	147	76.47059	62.18487	30.88235	56.53846	0.455882	0.588
3	53	78	25	156	77.94118	65.54622	36.76471	60	0.411765	0.624
4	51	54	25	130	75	45.37815	36.76471	50	0.382353	0.52
5	45	23	11	79	66.17647	19.32773	16.17647	30.38462	0.5	0.316
6	48	39	21	108	70.58824	32.77311	30.88235	41.53846	0.397059	0.432
7	40	29	12	81	58.82353	24.36975	17.64706	31.15385	0.411765	0.324
8	49	70	32	151	72.05882	58.82353	47.05882	58.07692	0.25	0.604
9	51	77	33	161	75	64.70588	48.52941	61.92308	0.264706	0.644
10	63	62	22	147	92.64706	52.10084	32.35294	56.53846	0.602941	0.588
11	36	23	13	72	52.94118	19.32773	19.11765	27.69231	0.338235	0.288
12	51	51	13	115	75	42.85714	19.11765	44.23077	0.558824	0.46
13	54	59	22	135	79.41176	49.57983	32.35294	51.92308	0.470588	0.54
14	31	41	17	89	45.58824	34.45378	25	34.23077	0.205882	0.356
15	46	62	13	121	67.64706	52.10084	19.11765	46.53846	0.485294	0.484
16	33	18	10	61	48.52941	15.12605	14.70588	23.46154	0.338235	0.244
17	47	32	13	92	69.11765	26.89076	19.11765	35.38462	0.5	0.368
18	54	63	14	131	79.41176	52.94118	20.58824	50.38462	0.588235	0.524
19	49	41	18	108	72.05882	34.45378	26.47059	41.53846	0.455882	0.432
20	54	49	14	117	79.41176	41.17647	20.58824	45	0.588235	0.468
21	44	29	13	86	64.70588	24.36975	19.11765	33.07692	0.455882	0.344
22	37	20	9	66	54.41176	16.80672	13.23529	25.38462	0.411765	0.264
23	43	40	17	100	63.23529	33.61345	25	38.46154	0.382353	0.4
24	51	73	31	155	75	61.34454	45.58824	59.61538	0.294118	0.62
25	50	46	14	110	73.52941	38.65546	20.58824	42.30769	0.529412	0.44
26	60	72	15	147	88.23529	60.5042	22.05882	56.53846	0.661765	0.588
27	31	23	9	63	45.58824	19.32773	13.23529	24.23077	0.323529	0.252
28	53	62	17	132	77.94118	52.10084	25	50.76923	0.529412	0.528
29	41	45	12	98	60.29412	37.81513	17.64706	37.69231	0.426471	0.392
30	58	51	20	129	85.29412	42.85714	29.41176	49.61538	0.558824	0.516
31	59	51	20	130	86.76471	42.85714	29.41176	50	0.573529	0.52
Mean	**47.87**	**47.97**	**17.74**	**113.58**	**70.398**	**40.309**	**26.091**	**43.685**	**0.443074**	**0.454323**

APPENDIX VIII

LEARNERS' SCORES ON EVEN AND ODD-NUMBERED ITEMS OF THE DEVELOPED TEST INSTRUMENT

	GRADE 9			GRADE 10			GRADE 11	
	EVEN	ODD		EVEN	ODD		EVEN	ODD
A91	82	81	A101	53	50	A111	67	69
A92	67	66	A102	27	31	A112	80	94
A93	47	47	A103	60	31	A113	80	44
A94	67	63	A104	53	50	A114	60	63
A95	50	53	A105	73	50	A115	67	56
A96	50	33	A106	73	50	A116	47	50
A97	31	27	A107	53	25	A117	27	13
A98	38	60	A108	53	56	A118	47	44
A99	19	33	A109	27	19	A119	67	69
A910	63	60	A1010	53	69	A1110	60	63
A911	38	13	A1011	27	44	A1111	67	75
A912	44	60	A1012	75	56	A1112	53	38
A913	44	33	A1013	53	38	A1113	40	38
A914	47	44	A1014	47	44	A1114	40	33
A915	67	66	A1015	40	44	A1115	47	49
A916	53	53	A1016	53	54	A1116	47	38
A917	49	50	A1017	53	59	A1117	47	31
A918	89	87	A1018	80	73	A1118	13	38
A919	53	51	A1019	63	56	A1119	80	50
A920	53	49	A1020	73	76	A1120	80	94
A921	58	56	A1021	47	50	A1121	60	94
A922	40	50	A1022	67	68	A1122	73	63
A923	93	44	A1023	33	38	A1123	80	69
A924	53	44	A1024	60	58	A1124	55	44
A925	53	31	A1025	57	63	A1125	53	56
A926	73	44	A1026	53	49	A1126	80	75
A927	47	38	A1027	60	55	A1127	73	44
A928	60	50	A1028	73	88	A1128	60	50
A929	53	31	A1029	53	57	A1129	47	56
B92	80	50	A1030	80	78	A1130	53	69
B93	27	38	B101	33	25	B111	60	50
B94	47	45	B102	37	32	B112	47	38
B95	49	47	B103	53	56	B113	60	50
B96	34	37	B104	53	50	B114	73	60
B97	23	25	B105	27	31	B115	53	56
B98	67	69	B106	60	31	B116	40	50
B99	40	38	B107	53	50	B117	67	54
B910	49	50	B108	73	50	B118	67	38
B911	46	48	B109	73	50	B119	40	41

B912	47	47	B1010	53	25	B1110	47	38
B913	27	28	B1011	53	56	B1111	60	69
B914	49	51	B1012	27	19	B1112	80	70
B915	31	34	B1013	53	69	B1113	47	44
B916	40	38	B1014	27	44	B1114	73	61
B917	34	38	B1015	75	56	B1115	57	63
B918	20	23	B1016	53	38	B1116	47	44
B919	43	46	B1017	60	61	B1117	53	63
B920	33	31	B1018	60	58	B1118	67	69
B921	47	43	B1019	73	71	B1119	53	44
B922	27	31	B1020	73	70	B1120	73	69
B923	53	48	B1021	53	50	B1121	53	63
B924	33	38	B1022	63	65	B1122	40	44
B925	33	31	B1023	73	66	B1123	40	31
B926	40	39	B1024	60	50	B1124	73	50
B927	33	25	B1025	33	31	B1125	47	38
B928	55	53	B1026	67	63	B1126	73	50
B929	27	24	B1027	43	46	B1127	53	50
B930	47	44	B1028	67	69	B1128	60	25
B931	34	38	B1029	73	72	B1129	47	38
B932	20	25	B1030	27	25	B1130	20	31
B933	47	41	C101	63	60	C111	80	73
C91	33	29	C102	63	51	C112	80	69
C92	40	38	C103	34	33	C113	67	75
C93	34	38	C104	56	53	C114	73	81
C94	33	29	C105	50	53	C115	87	69
C95	73	71	C106	56	54	C116	67	88
C96	45	50	C107	63	49	C117	60	63
C97	40	41	C108	50	53	C118	80	88
C98	53	56	C109	63	73	C119	47	38
C99	47	50	C1010	38	47	C1110	60	50
C910	40	38	C1011	40	44	C1111	73	50
C911	56	55	C1012	53	44	C1112	53	56
C912	44	47	C1013	53	44	C1113	40	50
C913	50	53	C1014	80	63	C1114	67	44
C914	50	33	C1015	73	44	C1115	67	38
C915	31	27	C1016	93	56	C1116	40	31
C916	38	60	C1017	27	50	C1117	47	38
C917	19	33	C1018	67	38	C1118	80	62
C918	63	60	C1019	33	38	C1119	87	75
C919	38	13	C1020	60	50	C1120	100	94
C920	44	60	C1021	73	56	C1121	67	81
C921	44	33	C1022	73	50	C1122	77	81
C922	44	40	C1023	53	50	C1123	53	63
C923	38	20	C1024	53	75	C1124	67	69
C924	31	7	C1025	44	67	C1125	53	44
C925	6	20	C1026	63	67	D111	73	69

C926	38	47	C1027	56	40	D112	53	63
C927	63	40	C1028	56	80	D113	40	44
C928	44	41	C1029	88	80	D114	40	31
D91	47	19	C1030	20	44	D115	73	50
D92	53	13	D101	33	31	D116	56	73
D93	27	31	D102	13	25	D117	31	47
D94	40	56	D103	20	19	D118	25	47
D95	47	25	D104	40	31	D119	50	33
D96	63	51	D105	33	38	D1110	25	13
D97	21	25	D106	60	19	D1111	50	40
D98	27	31	D107	20	19	D1112	38	13
D99	20	19	D108	40	31	D1113	56	73
D910	34	38	D109	33	38	D1114	50	60
D911	40	44	D1010	50	49	D1115	50	60
N = 100			**N= 100**			**N - 100**		
Pearson product	**0.683**		**r = 0.66953**			**r = 0.6806**		
S-B.P.F	**0.811**		**R = 0.8021**			**R = 0.810**		
Stdev	**15.4**	**14.7**		**16.8**	**15.7**		**16.2**	**17.9**

AVERAGE Reliability of the instrument = 0.80780

AVERAGE Standard deviation = 16.12

S-B.P.F = Reliability determined using the Spearman-Brown prophecy formula

APPENDIX IX

DATA USED TO CALCULATE THE READABILITY LEVEL OF THE INSTRUMENT.

KEY

Sample = Number of sampled item.

# of sentence = Number of sentences in the item.

# of words = Number of words per sentence.

# of syllables = number of syllables per word.

Ave. # of syllables = Average number of syllables per word.

ASL = Average sentence length.

ASW = Average number of syllables per word.

Sample	# of sentences	# of words	# of syllables	Ave # of syllables
Qn 2.	1	16	22	1.375
Qn 5	1	11	13	1.181818
Qn 7	1	14	21	1.5
"	2	16	23	1.4375
"	3	18	30	1.666667
"	4	15	18	1.2
"	5	14	18	1.285714
Qn 8	1	12	19	1.583333
"	2	16	22	1.375
"	3	20	28	1.4
"	4	30	20	0.666667
Qn 9	1	14	28	2
"	2	13	23	1.769231
"	3	5	7	1.4
Qn 11	1	11	16	1.454545
"	2	8	9	1.125
"	3	21	31	1.47619
Qn 13	1	20	28	1.4
"	2	9	15	1.666667
"	3	16	24	1.5
"	4	13	16	1.230769
Qn 15	1	19	24	1.263158
"	2	9	14	1.555556
Qn 18	1	17	21	1.235294
"	2	14	17	1.214286
"	3	21	27	1.285714
"	4	21	27	1.285714
"	5	25	35	1.4
Qn 21	1	27	32	1.185185
"	2	19	26	1.368421
Qn 23	1	10	16	1.6

"	2	8	10	1.25
"	3	11	16	1.454545
"	4	13	17	1.307692
"	5	17	28	1.647059
Qn 24	1	26	37	1.423077
"	2	21	38	1.809524
Qn 26	1	12	15	1.25
"	2	27	37	1.37037
"	3	17	24	1.411765
Qn 27	1	19	31	1.631579
"	2	10	18	1.8
Qn 30	1	11	19	1.727273
	ASL =	**15.95349**	**ASW =**	**1.422565**

APPENDIX X

DATA USED FOR THE CORRELATION OF TIPS AND DEVELOPED TEST SCORES

GRADE 9			**GRADE 10**			**GRADE 11**		
	D.T	**TIPS**		**D.T**	**TIPS**		**D. T**	**TIPS**
H91	74	64	H101	56	56	H111	73	52
H92	74	52	H102	68	68	H112	79	60
H93	47	44	H103	35	28	H113	47	44
H94	82	56	H104	85	56	H114	73	68
H95	65	36	H105	82	64	H115	85	72
H96	50	40	H106	65	64	H116	76	56
H97	50	40	H107	65	64	H117	79	88
H98	50	48	H108	65	52	H118	79	72
H99	74	60	H109	88	68	H119	82	72
H910	68	48	H1010	56	20	H1110	65	48
H911	79	60	H1011	71	52	H1111	82	56
H912	68	60	H1012	50	40	H1112	59	48
H913	76	60	H1013	62	36	H1113	68	64
H914	62	36	H1014	71	64	H1114	71	64
H915	62	48	H1015	41	12	H1115	50	48
H916	53	48	H1016	71	60	H1116	85	68
H917	59	48	H1017	38	48	H1117	65	76
H918	65	48	H1018	56	52	H1118	65	44
H919	68	48	H1019	65	52	H1119	73	64
H920	62	52	H1020	71	60	H1120	68	60
H921	59	44	H1021	67	44	H1121	73	60
H922	68	44	H1022	59	56	H1122	73	68
H923	65	56	H1023	85	52	H1123	71	64
H924	71	68	H1024	68	56	H1124	76	60
H925	76	40	H1025	68	40	H1125	79	68
H926	76	64	H1026	71	52	H1126	68	44
H927	56	28	H1027	56	43	H1127	73	60
H928	56	56	H1028	50	56	H1128	71	44
H929	68	68	H1029	56	64	H1129	79	64
H930	56	60	H1030	65	68	H1130	71	60
N = 30	r =	0.503	**N = 30**	r =	0.568	**N = 30**	r =	0.599

AVERAGE R = 0.5565793

Appendix XI

DISCRIMINATION AND DIFFICULTY INDICES FROM THE PILOT STUDY RESULTS.

KEY:
H = High scorers
L = Low scorers
Discrim = Discrimination index
Diff = Index of difficulty
Number of subjects = 150

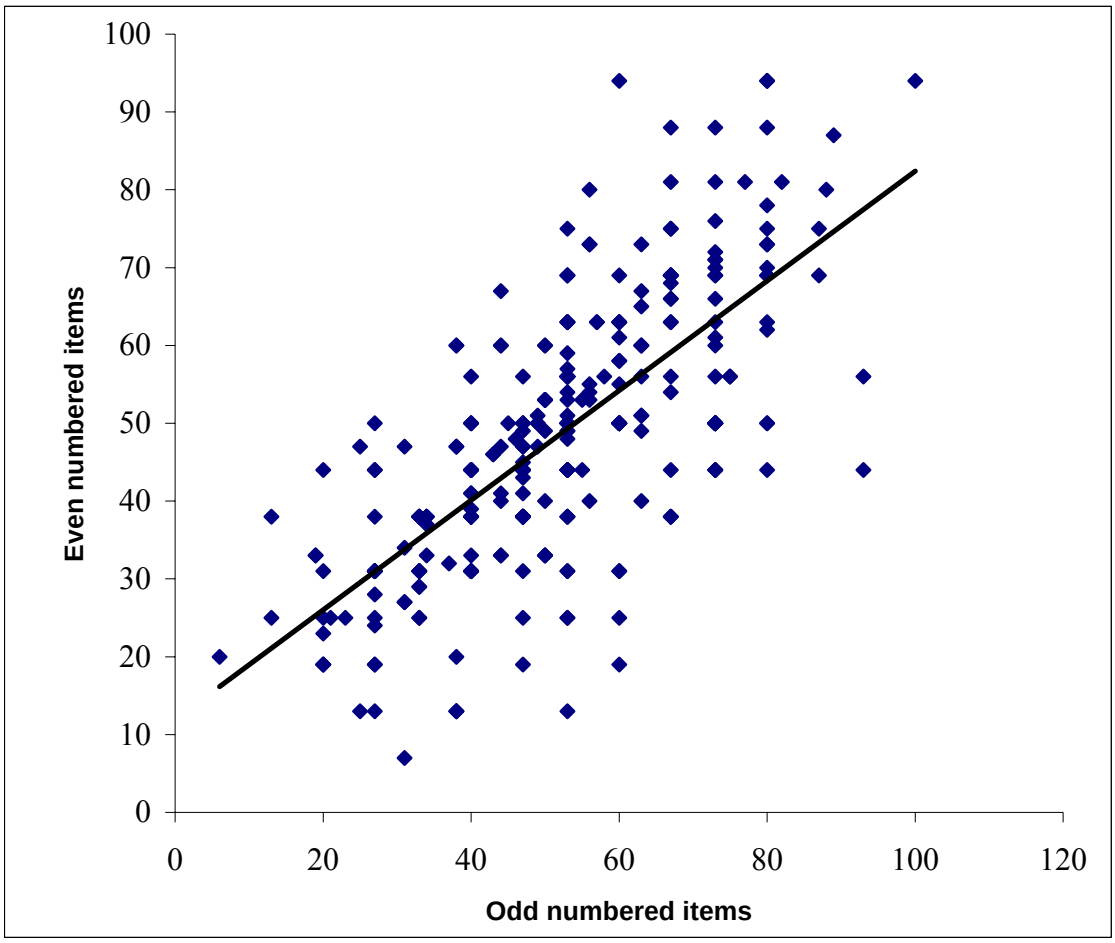
Item no.	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**
H	40	34	40	37	36	34	40	40	27	40	27	34	40	27	34	37	40	34	34	39	40	14	34	40	32	40	27	40	40
L	40	22	21	16	20	21	12	34	14	40	14	27	27	0	20	29	34	40	14	33	20	0	14	34	18	27	7	34	14
H - L	0	12	19	21	16	13	28	6	13	0	13	7	13	27	14	8	6	-6	20	6	20	14	20	6	14	13	20	6	26
Discrim	0	0.3	0.5	0.5	0.4	0.3	0.7	0.1	0.3	0	0.3	0.2	0.3	0.7	0.3	0.2	0.1	-1	0.5	0.1	0.5	0.3	0.5	0.1	0.3	0.3	0.5	0.1	0.6
H + L	80	56	61	53	56	55	52	74	41	80	41	61	67	27	54	66	74	74	48	72	60	14	48	74	50	67	34	74	54
Diff.	1	0.7	0.8	0.7	0.7	0.7	0.6	0.9	0.5	1	0.5	0.8	0.8	0.3	0.7	0.8	0.9	0.9	0.6	0.9	0.7	0.2	0.6	0.9	0.6	0.8	0.4	0.9	0.7
Item no.	**30**	**31**	**32**	**33**	**34**	**35**	**36**	**37**	**38**	**39**	**40**	**41**	**42**	**43**	**44**	**45**	**46**	**47**	**48**	**49**	**50**	**51**	**52**	**53**	**54**	**55**	**56**	**57**	**58**
H	40	40	27	40	34	38	34	40	40	34	40	40	34	40	27	34	34	27	36	27	40	40	14	34	40	34	40	35	31
L	34	20	7	20	40	18	20	14	34	14	40	34	27	27	0	20	20	14	16	14	34	20	0	14	34	20	27	21	25
H - L	6	20	20	20	-6	20	14	26	6	20	0	6	7	13	27	14	14	13	20	13	6	20	14	20	6	14	13	14	6
Discrim	0.1	0.5	0.5	0.5	-1	0.5	0.3	0.6	0.1	0.5	0	0.1	0.2	0.3	0.7	0.3	0.3	0.3	0.5	0.3	0.1	0.5	0.3	0.5	0.1	0.3	0.3	0.3	0.1
H + L	74	60	34	60	74	56	54	54	74	48	80	74	61	67	27	54	54	41	52	41	74	60	14	48	74	54	67	56	56
Diff.	0.9	0.7	0.4	0.7	0.9	0.7	0.7	0.7	0.9	0.6	1	0.9	0.8	0.8	0.3	0.7	0.7	0.5	0.6	0.5	0.9	0.7	0.2	0.6	0.9	0.7	0.8	0.7	0.7

Average discrimination index = 0.32

Average Index of difficult y =0 .722

APPENDIX XII

SCATTER DIAGRAM SHOWING THE RELATIONSHIP BETWEEN SCORES ON EVEN AND ODD NUMBERED ITEMS OF THE INSTRUMENT.



Correlation (after adjustment, using the Spearman – Brown Prophecy formula) R = 0.81